



Universität Hamburg
Fakultät für Mathematik,
Informatik und Naturwissenschaften
Zentrum für Bioinformatik

Bachelorarbeit

**Eine auf FISH Oracle 2
basierende Analyse von
Hochdurchsatz-Sequenz-Daten
aus dem TCGA-Projekt**

Sebastian Vincent Weigel

9weigel@informatik.uni-hamburg.de

Studiengang: Computing in Science mit Schwerpunkt Biochemie

Matr.-Nr.: 6126603

Erstgutachter Universität Hamburg:

Zweitgutachter Universität Hamburg:

Prof. Dr. Stefan Kurtz

PD Dr. Ronald Simon

Inhaltsverzeichnis

1 Einleitung	1
2 Grundlagen	3
2.1 Institute und <i>SNP-Microarrays</i>	3
2.2 Fish Oracle 2	5
2.3 fcd-Parser	8
2.4 Bildverarbeitung	9
3 Analyse und Ergebnisse	13
3.1 Chromosom 1	13
3.2 Chromosom 2	14
3.3 Chromosom 4	15
3.4 Chromosom 5	16
3.5 Chromosom 6	17
3.6 Chromosom 7	17
3.7 Chromosom 9	18
3.8 Chromosom 10	18
3.9 Chromosom 11	19
3.10 Chromosom 12	21
3.11 Chromosom 16	21
3.12 Chromosom 17	21
3.13 Chromosom 19	21
3.14 Chromosom 20	22
3.15 Chromosom X	24
3.16 Chromosom Y	24
3.17 Analyse der häufigsten Peaks	24
4 Diskussion	27
5 Ausblick	33
A Anhang	35
Literaturverzeichnis	77

1 Einleitung

Die vorliegende Arbeit beschäftigt sich mit den *Copy Number Variations* (CNVs, auch bekannt unter der Bezeichnung *Somatic Copy-Number Alterations*, SCNAs) von 18 verschiedenen Tumorentitäten. Die zugrunde liegenden Daten der CNVs wurden durch *SNP-Microarrays* gewonnen und beschreiben die vom Normalzustand abweichende Anzahl von bestimmten DNS-Abschnitten im Genom der einzelnen Patienten. Diese DNS-Abschnitte können codierende Abschnitte, also Gene, enthalten. Es wurden auch CNVs identifiziert, die vermutlich für geistige Behinderungen, Herzfehler oder Entwicklungsstörungen verantwortlich sind [8]. Auch in der Krebsforschung spielen CNVs eine immer größere Rolle. So haben einige Tumore spezifische CNVs, also charakteristische Muster in Genamplifikationen und Gendeletionen. Deshalb werden diese Daten auch mit in dem *The Cancer Genome Atlas* (TCGA) gesammelt. Diese Arbeit dokumentiert den Versuch, die untersuchten Tumorentitäten anhand ihrer CNVs zu gruppieren (*clustern*), die verwendeten Programme sowie die damit verbundenen Herausforderungen.

Das für diesen Zweck entwickelte Programm *GISTIC* (*Genome Identification of Significant Targets in Cancer*) arbeitet auf dem gesamten Genom [11]. Eine typische Ausgabe ist in Abbildung 2.3 dargestellt. Diese Abbildung ist [14] entnommen und zeigt Amplifikationen und Deletionen bei Harnblasenkrebs-Patienten.

Im Gegensatz zu *GISTIC* werden in dieser Arbeit die CNVs pro Chromosom betrachtet. Dies erlaubt einerseits eine höhere Auflösung der Daten, andererseits auch eine Gruppierung von Tumorentitäten nach ähnlichen Mustern auf den einzelnen Chromosomen. Dieser Ansatz wurde gewählt, nachdem sich in der frühen Phase der Arbeit zeigte, dass sich über CNVs über dem gesamten Genom auf dem vorliegenden Datensatz keine Gruppen bilden ließen.

Eine Gruppierung nach Chromosomen erlaubt möglicherweise auch Rückschlüsse auf gemeinsame Mechanismen in der Tumorentstehung der verschiedenen Entitäten.

Eine besondere Herausforderung ist die Visualisierung der Daten für Analysten. Eine Visualisierung der vormals nur tabellarisch vorliegenden Daten ist notwendig, damit Menschen in den Daten auf intuitive Art und Weise Muster erkennen können. Allerdings führen die 18 in dieser Arbeit betrachteten Tumorentitäten bei einer Aufschlüsselung in ihren CNVs je Chromosom bereits zu $18 \cdot 23 = 414$ Diagrammen.

Die zugrunde liegenden Daten stammen von dem *The Cancer Genome Atlas* (TCGA, siehe Kapitel 2.1) Projekt. Zunächst wurden diese Daten mit einem selbstgeschriebenen *JSON*-Parser (siehe Kapitel 2.3) in ein für *FISH Oracle 2* (siehe Kapitel 2.2) importierbares Format gebracht. Für jede Tumorentität wird dazu in *FISH Oracle 2* ein eigenes Projekt angelegt. Dabei werden die Datensätze auch zwischen der Kombination aus Institut und *SNP-*

Array-Platte unterschieden. Kapitel 3 enthält die Analyse ausgewählter Chromosomen. In Kapitel 2 werden die Ergebnisse diskutiert und die dabei aufgetretenen Probleme erläutert. Kapitel 5 bietet einen Ausblick auf weitere mögliche Forschungen auf diesem Gebiet.

2 Grundlagen

Die Kopiezahldaten (CNV-Daten) von Tumoren werden ermittelt, indem sowohl gesundes, als auch karzinöses Gewebe vom Patienten entnommen wird. Beide Proben werden mittels *SNP-Microarrays* analysiert und die Ergebnisse miteinander verglichen. Dabei kann festgestellt werden, ob in der Probe aus dem karzinösen Gewebe bestimmte Bereiche des Genoms vermehrt oder entfernt wurden. Wenn in der Tumorprobe DNS-Abschnitte entfernt wurden, spricht man hierbei von Deletion beziehungsweise Amplifikation, wenn DNS-Abschnitte vermehrt wurden. Die CNVs gewinnen in der molekularbiologischen Forschung immer mehr an Bedeutung, beispielsweise in der Genregulation. Nach jüngsten Analysen gibt es im menschlichen Genom knapp 28 000 CNVs [12].

Die hier eingesetzten Daten entstammen dem *TCGA* vom 13.06.2014. Das 2006 gestartete *TCGA*-Projekt hat es sich zur Aufgabe gemacht, eine Bibliothek mit Proben und prozessierten Daten von möglichst allen Tumorentitäten aufzubauen. Ziel ist eine Datenbank aller bekannter Krebsarten und ihrer Gendefekte zu erhalten. Des Weiteren werden vom *TCGA Research Network* regelmäßig auf diesen Daten basierende Publikationen veröffentlicht. Diese beschäftigen sich meist mit nur einer Tumorentität und nutzen neben den CNV-Daten noch weitere Analysemethoden, wie beispielsweise Mikro-RNA-Sequenzierung, DNA-Methylierung oder Exom-Sequenzierung[7].

2.1 Institute und *SNP-Microarrays*

Es gibt drei Institute im *TCGA*-Projekt, die eine CNV-Identifizierung im Genom mit *SNP-Microarrays* anbieten. Zwei Institute nutzen hierbei dieselben *SNP-Microarrays*, das dritte nutzt zwei verschiedene *SNP-Microarrays*. Eine Übersicht befindet sich in Tabelle 2.1.

Die Bereitstellung der Daten von verschiedenen Instituten und die Verwendung von verschiedenen *SNP-Microarrays* hat zur Folge, dass die Daten nicht einheitlich sind.

CPL	BGW6	WG6	HH1M	HHH5
Institut	Broad Institute (BI)[15]	The Genome Institute (WUSM)[4]	Hudson Alpha (HABI)[6]	
SNP-Microarray Hersteller	Genome-Wide Human SNP Array 6.0 Affymetrix[2]			Human1M-Duo Illumina[19]

Tabelle 2.1: Institute und *SNP-Microarrays*. Aus den Anfangsbuchstaben des Instituts und drei Zeichen der *SNP-Microarrays*-Platte ergeben sich die Kürzel (CPL) die zur Identifikation der Daten verwendet werden. Beispielsweise steht HH1M für Hudson Alpha Human1M-Duo.

CPL	Dateiendung	Datei-Header
BGW6	.hg18.seg.txt .nocnv_hg18.seg.txt .hg19.seg.txt .nocnv_hg19.seg.txt	ID chrom loc.start loc.end num.mark seg.mean
WG6	.segmented.dat	ID chrom loc.start loc.end num.mark seg.mean
HH1M & HHH5	.seg.txt .segnormal.txt .loh.txt	Barcode chrom loc.start loc.end mean

Tabelle 2.2: Nomenklatur der Quelldaten. Für die CPL-Kürzel siehe Tabelle 2.1. Übernommen von [3].

Datenformate

Die CNV-Daten (*SNP Array*) liegen beim *TCGA*-Projekt in 3 Leveln vor:

1. Rohdaten (pro Patient)
2. Unnormalisierte *SNP*-, *LOH*- und *CNV*-Daten (pro Probe)
3. Normalisierte *CNV*-Daten

In dieser Arbeit werden die Daten von Level 3 verwendet, also normalisierte *CNV*-Daten. Daten tieferen Levels sind auch nicht öffentlich zugänglich. Diese Daten liegen in spaltenbasierten Textdokumenten mit verschiedenen Dateiendungen vor. In Tabelle 2.2 ist eine Übersicht zu finden.

In den Dateien von *HH1M* und *HHH5* können auch Daten mehrerer Proben und Patienten vorliegen. Des Weiteren fehlt die *num.mark*-Spalte (siehe Tabelle 2.2).

Das *hg18* beziehungsweise *hg19* steht für die *Human Genome Version*. Dies ist die *UCSC* Nomenklatur. *hg19* entspricht dabei der *NCBI* Nomenklatur *GRCh37* (2009) und *hg18 NCBI36* (2006). Die *nocnv*-Daten beinhalten in diesem Fall keine Keimbahn-CNVs und sind für diese Art der Weiterverarbeitung zu präferieren [18]. Die *loh*-Daten zeigen nur einen kleinen Teil der gesamten CNVs, den Verlust der Heterozygotie (aus dem Englischen *loss*

of heterozygosity). In diesem Fall werden nur CNVs markiert, wenn bei einem Allel eine Deletion vorliegt. Die *segnormal*- und *seg*-Dateien enthalten die kompletten CNVs, mit dem Unterschied, dass die *segnormal*-Dateien normalisiert wurden. Es folgt eine Auflistung der Dateiendungen, die zur Weiterverarbeitung mit *FISH Oracle 2* verwendet wurden:

- .nocnv_hg19.seg.txt
- .segmented.dat
- .seg.txt

Die Daten lassen sich über eine Datenmatrix vom *TCGA Data Portal* [17] als Archiv lokal speichern. Eine Übersicht über die Tumorentitäten und deren am 13.06.2014 vorliegenden *SNP-Microarray*-Daten findet sich in der angehängten Tabelle A.1. Wie vom *TCGA*-Projekt vorgeschlagen, werden die Datensätze, die weit weniger als 100 Patienten beinhalten, hier nicht mit aufgenommen und katalogisiert. In der Tabelle A.1 sind diese Einträge rot hinterlegt.

In jedem *TCGA*-Archiv befindet sich eine *Manifest*-Datei. Auch diese liegt als ein spaltenbasiertes Textdokument vor. Es beinhaltet alle zu ihrem Archiv dazugehörigen Dateien. Jede Datei steht dabei in einer neuen Zeile. Jede Zeile enthält dabei die der Tabelle 2.3 zu entnehmenden Informationen.

Spalte	Bezeichnung	Bedeutung	Beispiel
1	Platform Type	Materialtyp	CNV (<i>SNP Array</i>)
2	Center	Institutskürzel	BI
3	Platform	Arrayplatte	<i>Genome_Wide_SNP_6</i>
4	Level	Datenlevel	3
5	Sample	Sample-ID	TCGA-32-1991-10
6	Barcode	TCGA-Barcode	TCGA-32-1991-10C-01D-1224-01
7	File Name	Dateiname	*.hg19.seg.txt

Tabelle 2.3: Format der *TCGA-Manifest*-Dateien. Da der Dateiname für diese Arbeit keine Relevanz hat, wurde er mit einem Sternchen (*) ersetzt.

Der Aufbau des *TCGA*-Barcodes (Spalte 6) besteht aus sieben beziehungsweise neun Abschnitten. Die Bedeutung der IDs sind der Tabelle 2.4 zu entnehmen.

2.2 Fish Oracle 2

FISH Oracle 2 ist eine Software zur einheitlichen visuellen Darstellung von Genomdaten in der Krebsforschung [9]. Ein *FISH Oracle 2* Web-Server wurde zu demonstrativen

¹100-120 mg

Bezeichnung	Bedeutung	Beispiel	Erklärung
Project	Projektname	TCGA	
TSS	Tumorentität Probenquelle	32 St. Joseph's Hospital (AZ)	GBM
Participant	Patient	1991	Patient mit der ID 1991
Sample	Probentyp	10	Normale Zellen aus Blut
Vial	Probennummer	C	Probe 3 von 10 Alphabetische Nummerierung
Portion	Portion ¹ der Probe	01	Portion 01 von 10C
Analyte	Analytkürzel	D	DNA
Plate	Nummer der 96-well-Platte	1224	1224. Array-Platte der Portion
Center	Nummer des Instituts	01	BI

Tabelle 2.4: Der Aufbau des TCGA-Barcodes an folgendem Beispiel:

TCGA-32-1991-10C-01D-1224-01

Übernommen von [5].

Zwecken unter [16] eingerichtet. Die Funktionalität beschränkt sich allerdings nicht nur auf die hier verwendete Visualisierung von *SNP-Microarray*-Daten. Mit dem Programm ist es möglich über eine Web-Oberfläche Grafiken, wie in Abbildung 2.1 zu generieren. Diese können in vier verschiedenen Formaten (*pdf*, *ps*, *svg* und *png*) exportiert und lokal gespeichert werden. In dieser Arbeit werden die Vektorgrafiken *svg* verwendet.

In *FISH Oracle 2* werden für CNV-Daten spaltenbasierte Textdokumente importiert. Die jeweilige Datei enthält die Daten für einen Patienten und muss sich aus folgende Spalten zusammensetzen [10]:

1. Chromosome column (“chrom”)
2. Start position (“loc.start”)
3. End position (“loc.end”)
4. Number of markers (“num.mark”)
5. Segment mean intensity value (“seg.mean”)

Es treten also zwei Probleme beim direkten Import der *TCGA-Level-3-Dateien* in *FISH Oracle 2* auf:

1. Die *num.mark*-Spalte bei den Dateien von *HH1M* und *HHH5* fehlen.
2. Die Dateien von *HH1M* und *HHH5* können Daten von mehreren Patienten enthalten.

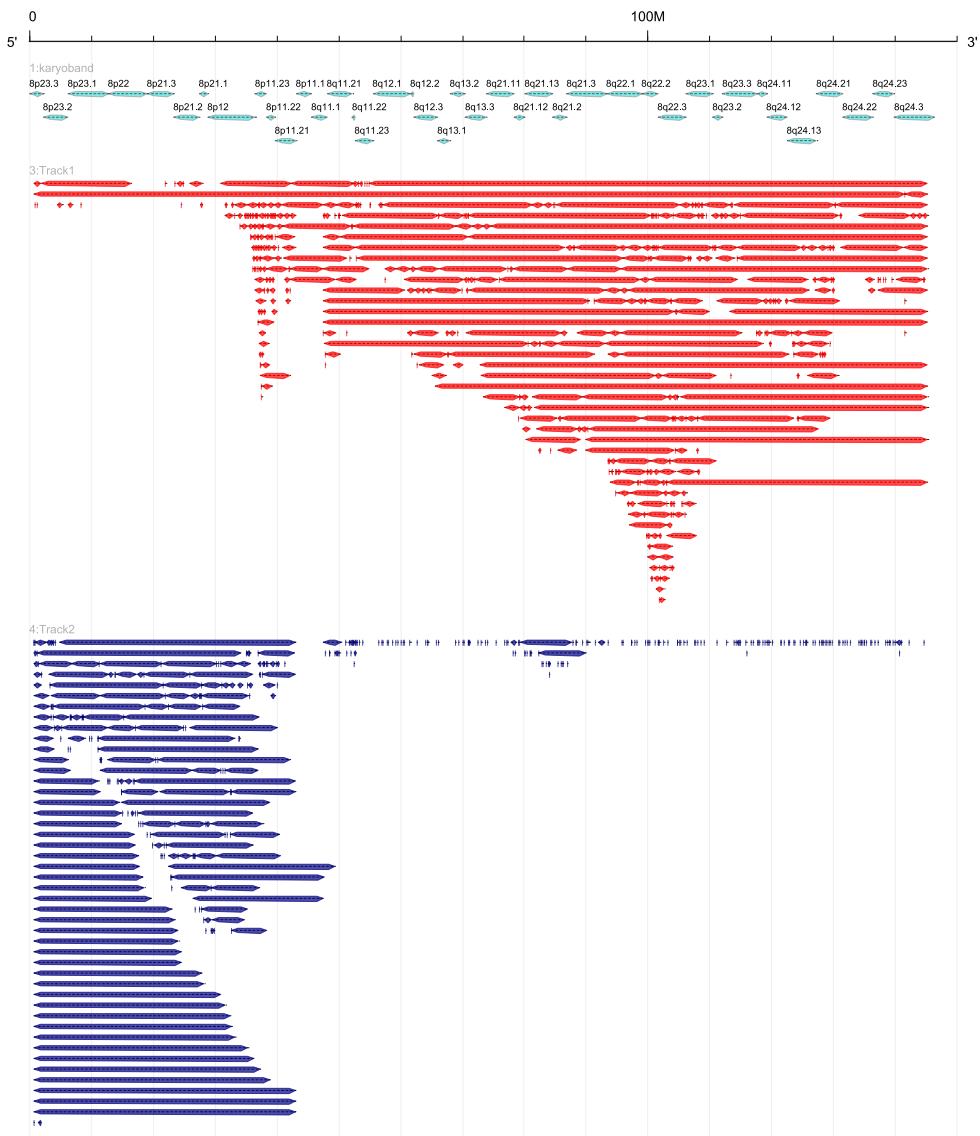


Abbildung 2.1: *SVG FISH Oracle* Ausgabe: Chromosom 8 von Harnblasenkrebs-Patienten (BLCA) vom BGW6. Track 1 ist das Chromosomenband (karyoband), welches die einzelnen Chromosomabschnitte zeigt. Track 3 zeigt die Amplifikationen in rot für das gesamte Chromosom 8. Es sind zwei deutliche (38 Mbp, 8p11.23; 103 Mbp, 8q22.2/8q22.3) und ein schwacher Peak (48 Mbp, 8q11.1) zu erkennen. Track 4 beinhaltet die Deletionen in blau für das gesamte achte Chromosom. Der p-Arm ist bei BLCA-Patienten im achten Chromosom stark von Deletionen betroffen, wohingegen die Amplifikationen sich abgesehen vom Peak im Abschnitt 8p11.23 eher im q-Arm befinden.

2.3 *focd*-Parser

Zur Lösung des in Abschnitt 2.3 beschriebenen Problems habe ich einen Parser in der Skriptsprache Perl geschrieben. Der *focd*-Parser wird mit mindestens vier Parametern über die Kommandozeile gestartet. Die Parameter können folgender Auflistung entnommen werden:

1. Projekt ID, ab der mit dem Import begonnen werden soll.
2. Quellpfad, in denen die entpackten TCGA-Archive mit ihrer vorgegebenen Ordnerstruktur liegen.
3. Zielpfad, wo die *focd*-Dateien abgespeichert werden sollen.
4. Institut und *SNP-Microarrays* von denen die Daten geparsst werden sollen:
 - 1) BI: Genome Wide SNP 6 (.nocnv_hg19.seg.txt)
 - 2) WUSM: Genome Wide SNP 6 (segmented.dat)
 - 3) HAIB: HumanHap550 (.seg.txt)
 - 4) HAIB: Human1MDuo (.seg.txt)

Zunächst werden alle Ordner nach *TCGA-Manifest*-Dateien durchsucht. Die für die Weiterverarbeitung der TCGA-Dateien wichtigen Daten befinden sich in den Spalten 1, 2, 3, 6 und 7 (siehe Tabelle 2.3). Befinden sich in einer Datei mehrere Patienten, sind ihre Barcodes mit einem Schrägstrich (/) getrennt.

Mit Hilfe eines Arrays in einer *Hash*-Tabelle werden die Daten pro Patient in eine neu erstellte *focd*-Datei geschrieben. Die Hash-Tabelle enthält die Dateinamen (Spalte 7) als Schlüssel (*key*). Der zugehörige Wert (*value*) ist ein Array. Die ersten beiden Einträge des Arrays sind das Institut (*center*) aus Spalte 2 und das *SNP-Microarray* (*platform*) aus Spalte 3. Im Anschluss folgt das Patienten-Array aus Spalte 6. Um auszuschließen, dass von einem Patienten von demselben Institut mit demselben *SNP-Microarray* mehrere Dateien vorliegen, wird der *TCGA*-Barcode um 17 Zeichen gekürzt. Somit besteht der Code, welcher Bestandteil des Namens der neu erstellten *focd*-Datei wird, nur noch aus folgenden IDs (Tabelle 2.4):

- TSS (Tumorentität)
- Participant (Patient)

Nun können die doppelten Patientendaten entfernt werden. Deutlich wird dies an folgendem Beispiel: Bei der Tumorentität GBM liegt eine Datei mit den Daten von der Probe mit dem *TCGA*-Barcode *TCGA-13-1489-02A-01D-0805-06* und eine Datei mit den Daten der Probe mit dem *TCGA*-Barcode *TCGA-13-1489-01A-01D-0474-06* vor. Die Barcodes haben im dritten Abschnitt dieselbe Nummer und sind somit ein Patient. Ohne die Prüfung mit den gekürzten Barcodes würden also einige CNV-Daten in mehrfacher Ausführung

vorliegen. Dies könnte zu einem verfälschten Ergebnis führen.

Des Weiteren werden nur Patienten-Daten in die *Hash*-Tabelle übernommen, die aus Tumorgewebe stammen. Also nur Dateien deren TCGA-Barcode an der Probentyp-Stelle (siehe Tabelle 2.3) eine 1 als erste Ziffer haben. Die somit noch in der *Hash*-Tabelle stehenden Dateien werden zunächst geöffnet und zeilenweise in ein Array eingelesen. Nun werden diese in die neu erstellte, zu ihrem Barcode passende *focd*-Datei übertragen. Stehen in einer Quelldatei also mehrere Proben, so werden die Zeilen den Proben nach auf die jeweils zugehörige *focd*-Datei verteilt. Hierbei steht der jeweilige Barcode in der ersten Spalte jeder Zeile und kann somit dem richtigen Patienten zugeordnet werden.

Der *focd*-Dateiname ist dabei wie folgt aufgebaut:

Tumorentitätskürzel_TSS-Participant_CPL.focd (Beispiel: GBM_32-1991_BGW6.focd).

Redo-Dateien beinhalten die Daten von wiederholten Versuchen. Dabei werden sie mit einem *_redo* gekennzeichnet und in einem extra Unterordner abgelegt. Sie wurden in dieser Arbeit allerdings nicht untersucht.

Da *FISH Oracle 2* die gängige wissenschaftliche Potenzschreibweise (1e+05 für 100.000) nicht verarbeiten kann, werden diese von dem *focd-Parser* in eine Dezimalzahl umgewandelt. Zusätzlich wird ein Bash-Skript und für jedes Archiv eine Manifestdatei erzeugt. Diese Kombination ermöglicht einen einfacheren und stabileren Datenimport in *FISH Oracle 2*.

2.4 Bildverarbeitung

Die für die Katalogisierung verwendeten Grafiken wurden für je ein *FISH Oracle 2* Projekt und pro Chromosom angelegt. Hierfür wurden zwei *Tracks* angelegt. Einer für die Deletion mit der Farbe blau und der zweite für die Amplifikation mit der Farbe rot. Der *intensity*-Wert wurde auf 0.5 festgelegt. Der Globale Schwellwert (*Global Segment Threshold*) wurde deaktiviert, damit jeder *Track* seinen eigenen *intensity*-Wert bekommt: Der Amplifikation-*Track* mit 0.5 und der Deletion-*Track* mit -0.5.

FISH Oracle 2 benutzt *AnnotationSketch* aus dem Paket *GenomeTools* zur Visualisierung. Um die Lesbarkeit zu erhöhen, wurden die von *FISH Oracle 2* erzeugten Grafiken mit Hilfe eines Python Skripts modifiziert. Für die Modifikation bot sich die *svg*-Ausgabe an, da es sich hierbei um ein Vektorgrafikformat auf Basis von *XML* handelt und sich die Veränderung automatisch auf alle exportierten Grafiken anwenden lies.

Das Skript leistet folgendes. Da die erzeugten Grafiken leider wenig Struktur aufweisen, dienen die grauen Beschriftungen der Tracks als Grenze zwischen den Tracks, die dann gruppiert werden. Die Deletionen (blau) werden am karyoband gespiegelt. Dies wird mit dem Attribut *transform="scale(1, -1)"* in der Gruppe der Deletionen realisiert. Anschließend wird das gesamte Bild mit dem Attribut *transform="rotate(90, 0,*

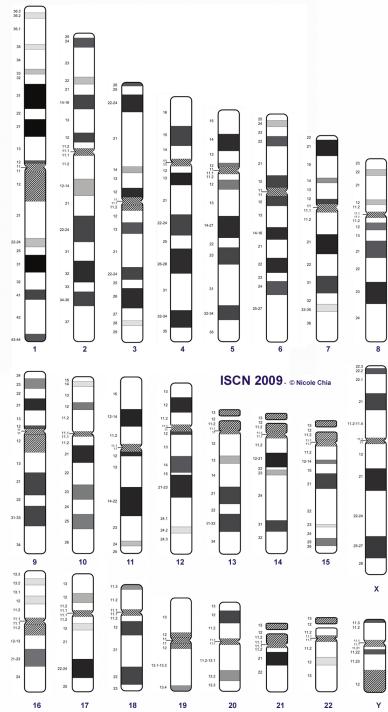


Abbildung 2.2: Klassische Karyogramme.

Quelle: Atlas of Genetics and Cytogenetics in Oncology and Haematology [1].

0) " um 90 Grad gedreht. Des Weiteren müssen im Header der *svg*-Datei noch die Höhe, Breite und das Sichtfeld angepasst werden. In Anlehnung an andere Karyogramme (Abbildung 2.2 und Abbildung 2.3) entsteht so eine gute Übersicht über Amplifikationen und Deletionen in einem Chromosom (Abbildung 2.4).

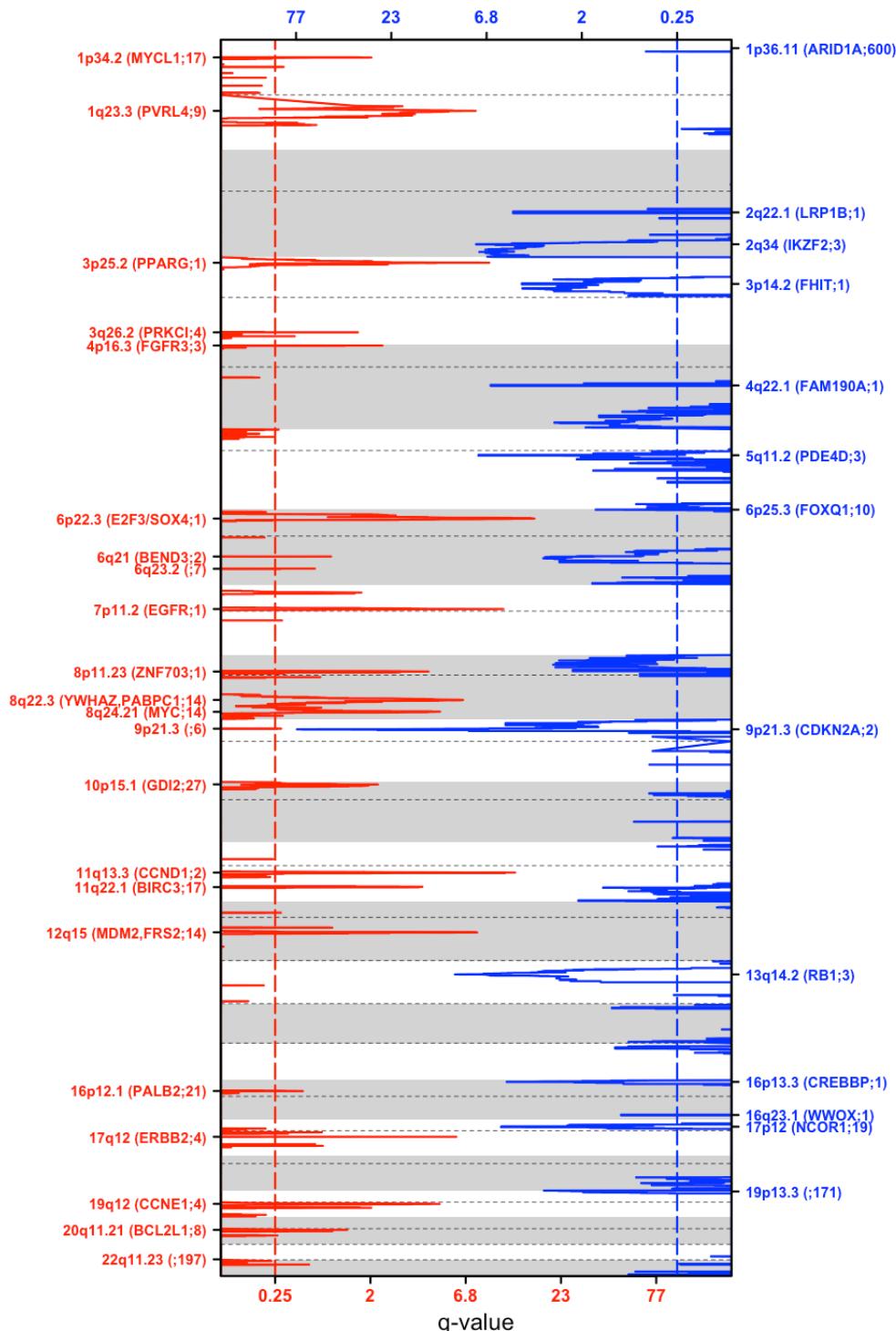


Abbildung 2.3: *GISTIC* Ausgabe vom Genom bei Harnblasenkrebs-Patienten (BLCA). Links sind in rot die Amplifikationen und rechts in blau die Deletionen dargestellt. Diese Darstellung ist wesentlich größer als die *FISH Oracle* Ausgabe, umfasst dafür aber das gesamte Genom. Für das achte Chromosom werden drei Amplifikationspeaks aufgeführt. 8p11.23 und 8q22.3 sind in der *FISH Oracle* Ausgabe (Abbildung 2.4) ebenfalls deutlich zu erkennen. 8q24.21 ist durch Überlagerungen der Amplifikationen nur schwach beziehungsweise schwierig zu identifizieren. Quelle: [14].

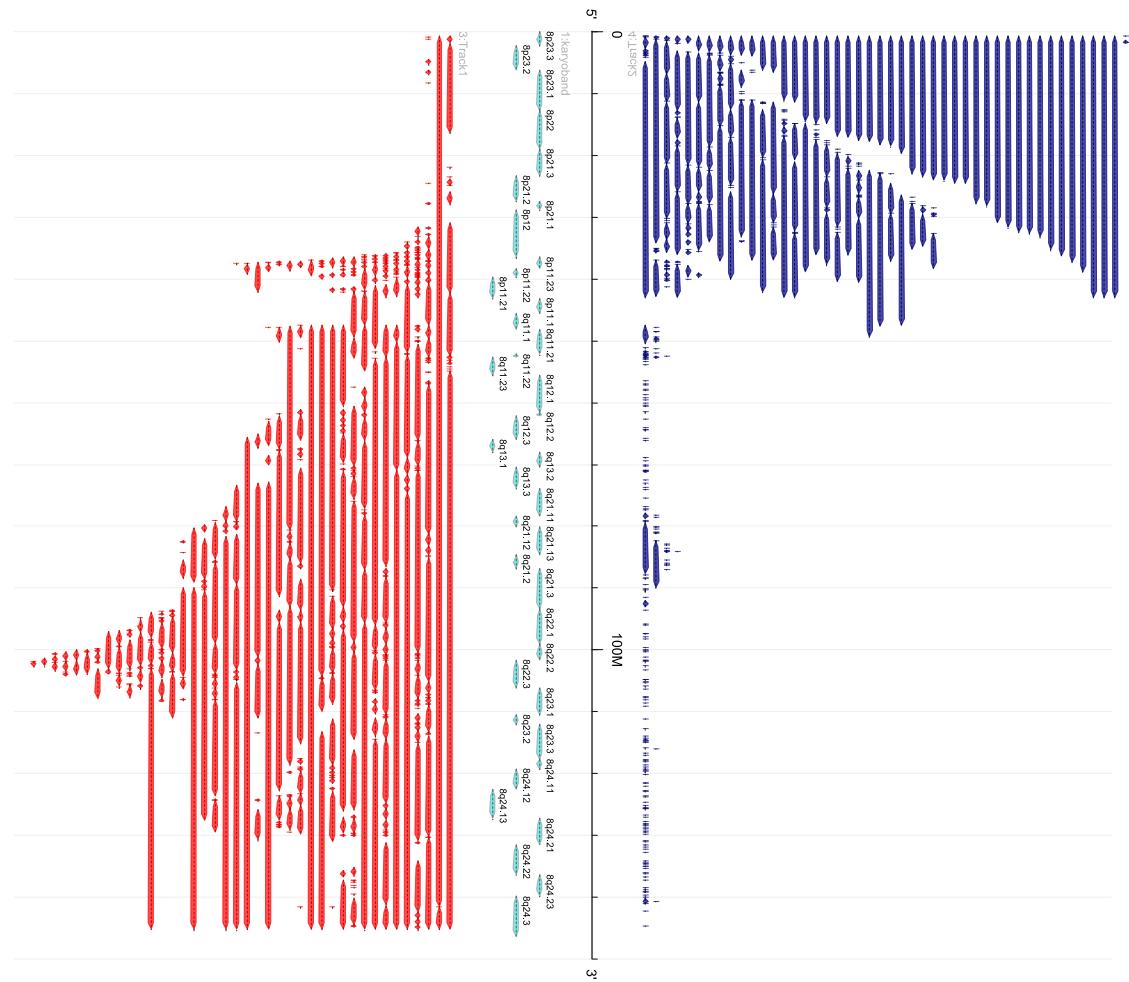


Abbildung 2.4: Modifizierte *SVG FISH Oracle* Ausgabe: Chromosom 8 von Harnblasenkrebs-Patienten (BLCA) vom BGW6. Auf der linken Seite der Grafik liegt nun der Track 3 mit den Amplifikationen in rot. In der Mitte befinden sich nun das Chromosomenband, sowie die Skala. Auf der rechten Seite steht der Track 4 mit den Deletionen in blau. Vergleiche hierzu Abbildung 2.1.

3 Analyse und Ergebnisse

Für jedes Chromosomen wurden die modifizierten *FISH Oracle*-Ausgaben der 18 verschiedenen Tumorentitäten miteinander verglichen. Hierbei wurden soweit möglich anhand der Muster der Deletionen und Amplifikationen auf den einzelnen Chromosomen Gruppen erstellt.

Besondere Muster bei den Entitäten werden zusätzlich beschrieben und hervorgehoben. Des Weiteren werden die Länge und die Position des Centromers mit angegeben. Die Positionen werden dabei in Mega-Basenpaaren (*Mbp*) angegeben. Das Centromer ist der Punkt im Chromosom, wo die Chromosomarme p und q aufeinander treffen. Der p-Arm ist dabei immer der kürzere. Die Arme sind zusätzlich in Regionen, diese in Bänder und diese wiederum in Subbänder aufgeteilt. Die Abschnitte steigen in ihrer Nummerierung vom Zentrum nach außen. Zu jedem Chromosom wurde zusätzlich eine Tabelle für auffällige Amplifikations- beziehungsweise Deletionspeaks erstellt, sofern welche vorhanden waren. Auffällig bedeutet in diesem Fall, dass die Peaks in mehreren Entitäten gut sichtbar ausgeprägt sind. Sehr stark ausgeprägte Peaks, die bei nur einer oder wenigen Entitäten vorkommen, werden ebenfalls mit aufgelistet. Hierbei sind die ungefähre Position, der Abschnitt und die Entitäten, die diesen Peak aufweisen, angegeben. Beinhaltet der Peak mehrere Abschnitte, so werden diese durch einen Schrägstrich getrennt. Bei den Peaks wird in gut erkennbar (*g*) und schwach ausgeprägt (*s*) unterschieden. Die am stärksten ausgeprägten Peaks bei einer Entität werden fett hervorgehoben.

Bei drei Entitäten (GBM, LAML und OV) wurden zwei verschiedene *SNP-Microarrays* eingesetzt. Sehr oft haben die doppelten Entitäten ein ähnliches Muster aufgewiesen und hatten dieselben Peaks, wenn auch nicht immer in der gleichen Stärke. Dabei sind die Peaks bei den Hudson Alpha Proben meist deutlicher erkennbar. Die Proben vom Broad Institut haben allerdings ein höhere Intensität. Wenn es in einem dieser Fälle bei einem Chromosom Unterschiede gab, werden diese extra erwähnt.

In dieser schriftlichen Ausarbeitung werde ich nur auf eine Auswahl an Chromosomen eingehen. Bei den in dieser Arbeit nicht behandelten Chromosomen waren keine Deletions- oder Amplifikationspeaks auffällig, sprich in mehreren Tumorentitäten gut sichtbar ausgeprägt.

3.1 Chromosom 1

Das erste Chromosom hat 249 250 621 Basenpaare (*bp*). Das Centromer befindet sich ungefähr bei 125 *Mbp*.

Bei LAML ist es beim 1. Chromosom kaum zu Deletionen und Amplifikationen gekommen. Bei THCA haben wir ein, bis auf die Deletionen ab 150 Mbp im q-Arm, ein ähnliches Bild. Deshalb werden diese beiden Entitäten bei den weiteren Erläuterungen nicht mehr mit einbezogen. Der p-Arm ist bei allen Entitäten stärker deletiert als der q-Arm und der q-Arm weist mehr Amplifikationen auf als der p-Arm. Das Muster von KIRP, KIRC und PRAD ist sowohl bei den Amplifikationen, als auch bei den Deletionen ein sehr ähnliches. Bei PRAD wurden allerdings mehr kurze Abschnitte deletiert. LGG ist die einzige Entität mit einer sehr starken Deletion des p-Arms. Auch wenn der p-Arm bei GBM nicht so stark deletiert wurde, wie bei LGG, so zeigt sich bei den beiden Entitäten ein sehr ähnliches Bild. Bei UCEC und BRCA sind beispielsweise sehr viele Amplifikationen am q-Arm entstanden. Ein sehr ähnliches Muster haben dabei HNSC, SKCM, LUSC, LUAD, OV STAD, COAD und READ. BLCA wird ebenfalls zu der Gruppe gezählt, auch wenn es die einzige Entität mit einem starken Peak um 160 Mbp besitzt. Ansonsten weisen alle 11 Entitäten sehr ähnliche Peaks auf (Tabelle 3.1).

Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
6	1p36	-	s	g	g	-	-	-	-	g	s	s	g	-	-	g	s	-	s
16	1p36	-	g	-	-	-	-	-	-	-	s	s	s	-	s	s	-	-	-
21	1p36	-	s	s	g	-	-	-	-	-	-	-	g	-	-	s	s	-	g
28	1p36	g	s	g	s	-	-	-	-	-	-	s	-	s	s	s	-	s	
150	1q21	g	s	-	-	s	-	-	-	-	g	s	g	-	s	g	g	-	g
161	1q23	g	g	-	-	-	-	-	-	-	s	g	-	-	s	-	s	-	s
206	1q32	s	g	-	g	-	-	-	-	g	s	s	-	-	s	g	-	-	g

Tabelle 3.1: Amplifikations- und Deletionspeaks vom ersten Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.2 Chromosom 2

Das zweite Chromosom hat 243 199 373 bp. Das Centromer liegt etwa bei 93.3 Mbp.

Auch beim zweiten Chromosom sind bei LAML und THCA kaum Deletionen und Amplifikationen vorhanden. Bei KIRP und READ ist dies ebenfalls der Fall. Am auffälligsten ist der Deletionspeak bei ungefähr 141 Mbp. Dieser ist bei 13 von den restlichen 14 Entitäten vorhanden (vergleiche Tabelle 3.2). Auch im Muster ähneln sich diese 14 Entitäten. Allerdings sind bei BRCA, OV, PRAD, COAD, GBM und BRCA im Bereich zwischen 100 Mbp und 170 Mbp ein erhöhtes Deletionsvorkommen (43%) und bei BRCA, OV, KIRC, COAD, GBM, BLCA, LGG, SKCM, UCEC, LUSC, LUAD und HNSC ein An-

stieg der Deletionsintensität ab 200 Mbp (86%) vorhanden. Die Amplifikationsintensität ist bei PRAD, KIRC, SKCM, LGG und COAD sehr ausgeglichen und es sind kaum Peaks erkennbar. Bei den restlichen Entitäten (BRCA, OV, BLCA, STAD, HNSC, LUAD, LUSC, UCEC und GBM) ist die Intensität im p-Arm ein wenig stärker als im q-Arm. Außerdem gibt es noch zwei Amplifikationspeaks im q-Arm, die bei fast allen Entitäten vorkommen. Ein Peak befindet sich im Bereich um die 180 Mbp (67%) und der andere um die 190 Mbp (56%). Alle zuletzt genannten Entitäten weisen aber mindestens einen von diesen beiden Peaks auf, mit Ausnahme von LUAD. Allerdings besitzt LUAD eine vermehrtes Aufkommen an Deletionen in diesem Bereich (170 Mbp - 200 Mbp).

Position (Mbp)	Chromosom-abschnitt																		
		BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
141	2q22	g	s	s	g	g	s	-	-	-	s	g	s	s	-	s	-	-	g
178	2q31	s	s	-	-	g	-	-	-	-	-	g	g	-	-	-	-	-	s
189	2q31	s	g	-	-	-	-	-	-	-	-	g	-	-	-	s	-	s	

Tabelle 3.2: Amplifikations- und Deletionspeaks vom zweiten Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.3 Chromosom 4

Das vierte Chromosom hat 191 154 276 bp. Das Centromer befindet sich in etwa an Position 50.4.

Bei der Deletion lassen sich die Entitäten in vier Gruppen aufteilen. Die Gruppen ergeben sich aus dem Deletionsmuster der einzelnen Grafiken. Bei Gruppe 1 gibt es am q- und am p-Ende ein vermehrtes Auftreten von Deletionen. Zu dieser Gruppe zählen BRCA, LUAD, UCEC und LUSC. Gruppe 2 besteht aus Entitäten, deren Deletionsmuster um die 90 Mbp ein vermehrtes Auftreten von Deletionen aufweisen. Die Deletionsintensität am q-Arm und p-Arm sind dabei nicht ganz einheitlich. Dieser Gruppe gehören OV, BLCA, COAD, HNSC, READ, PRAD und LAML an. Bei Gruppe 3 nehmen die Deletionen zum 3'-Ende, also zum q-Arm hin, zu. Zu dieser Gruppe zählen GBM, KIRC, LGG, STAD und THCA. Gruppe 4 enthält Entitäten, deren Deletionsmuster sehr geradlinig sind. Hierbei gibt es keine Auffälligkeiten und die Deletionen sind gleichmäßig über das gesamte Chromosom verteilt. Hierzu zählen KIRP und SKCM.

Nach der Amplifikation lassen sich die Tumorentitäten in drei Gruppen aufteilen. Gruppe 1 besitzt im Bereich um die 55 Mbp einen Peak. Hierzu zählen COAD, GBM, KIRC, LGG, LUSC und SKCM. Gruppe 2 hingegen besitzt einen ausgeprägten Peak auf der

Höhe von 75 Mbp und beinhaltet die Tumorentitäten OV, PRAD, UCEC, STAD, LUAD und BRCA. Gruppe 3 besteht auch den Tumorentitäten die sich keiner der ersten beiden Gruppen zuteilen ließen.

Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
55	4q12	-	-	s	g	s	s	-	-	g	s	g	s	-	s	s	-	-	s
178	4q34	s	s	s	g	g	-	-	-	s	-	-	g	-	-	g	s	s	g

Tabelle 3.3: Amplifikations- und Deletionspeaks vom vierten Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.4 Chromosom 5

Das fünfte Chromosom hat 180 915 260 bp. Das Centromer befindet sich ungefähr bei 48.4 Mbp.

Die Entitäten lassen sich anhand ihres Deletionsmusters in fünf Gruppen sortieren. THCA ist die einzige Entität, welche nur im p-Arm des Chromosoms deletiert und bildet eine eigene Gruppe. In Gruppe 2 sind die Deletionen gleichmäßig über das Chromosom verteilt, einschließlich des p-Arms. Dazu zählen SKCM, LGG, GBM, KIRP. Die Entitäten aus den restlichen drei Gruppen haben eine Gemeinsamkeit: Am p-Arm des Chromosoms ist nur eine sehr geringe Intensität an Deletionen. Gruppe 3 enthält die Entitäten LAML und HNSC. Diese haben ein ähnliches Deletionsmuster wie die Entitäten aus Gruppe 2, unterscheiden sich allerdings durch eine höhere Deletionsintensität beim p-Arm. Die Gruppen 4 und 5 zeichnen sich durch eine starke Ansammlung an Deletionen auf der ungefähren Position 60 und 110 aus. Gruppe 4 (COAD und READ) hat eine stärkere Intensität von Deletionen um die Position 110 000 000. Die restlichen Entitäten, OV, BLCA, BRCA, KIRC, LUAD, LUSC, PRAD, STAD und UCEC, gehören zu Gruppe 5. Im Gegensatz zu Gruppe 4 kommt eine stärkere Intensität von Deletionen um die Position 60 vor. In Tabelle 3.4 steht die Position und der Chromosomabschnitt eines Deletionspeaks, die bei einigen Entitäten sehr stark ausgeprägt sind und in mehreren Entitäten vorkommen.

Aus den Amplifikationsmustern ergeben sich ebenfalls fünf Gruppen. Gruppe 1 zeichnet sich durch ein hohe Amplifikationsrate im p-Arm und durch sehr geringe bis keine Amplifikationen im q-Arm aus. Zu dieser Gruppe zählen COAD, HNSC und LUSC. Gruppe 2 hat ebenfalls ein erhöhtes Vorkommen von Amplifikationen im p-arm. Im Gegensatz zu Gruppe 1 gibt es aber auch ein Amplifikationsvorkommen im q-Arm. Zu dieser Gruppe

gehören BLCA, PRAD, READ, SKCM und STAD. Gruppe 3 hat ein ähnliches Muster wie Gruppe 2. Hier ist lediglich das Ende des q-Arms noch weiter erhöht als beim restlichen q-Arm. Diese Gruppe besteht aus BRCA, LUAD, OV und UCEC. Gruppe 4 (KIRP, LGG, THCA und KIRC) weist ein gleichmäßig verteiltes Muster auf. Bei der fünften Gruppe sind kaum bis keine Amplifikationen vorhanden. Nur bei GBM, neben LAML das einzige Mitglied von Gruppe 5, ist ein kleiner Peak zu erkennen. Die Amplifikationspeaks sind der Tabelle 3.4 zu entnehmen. Insgesamt gab es für die Amplifikationen beim fünften Chromosom nur zwei auffällige Peaks.

Position (Mbp)	Chromosomabschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
59	5q11/q12	g	-	g	-	g	-	s	-	-	s	g	g	-	g	s	g	-	g

Tabelle 3.4: Deletionspeaks vom fünften Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.5 Chromosom 6

Das sechste Chromosom hat 171 115 067 bp. Das Centromer liegt circa bei 61 Mbp.

Position (Mbp)	Chromosomabschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
108	6q21	g	g	-	-	-	-	-	-	-	-	-	g	s	-	-	g	-	s

Tabelle 3.5: Amplifikationspeaks vom sechsten Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.6 Chromosom 7

Das siebte Chromosom hat 159 138 663 bp. Das Centromer befindet sich ungefähr bei 59.9 Mbp.

Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
56	7p11	g	g	s	g	g	-	-	-	g	g	g	-	-	-	-	g	-	g

Tabelle 3.6: Amplifikationspeaks vom siebten Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.7 Chromosom 9

Das neunte Chromosom hat 141 213 431 bp. Das Centromer befindet sich ungefähr bei 49 Mbp.

Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
9	9p11	s	g	-	s	g	g	-	-	-	g	g	g	-	-	g	g	-	s
22	9p21	g	g	g	g	g	g	-	-	g	g	g	-	-	g	g	-	s	
35	9p13	s	g	s	g	g	-	s	-	-	-	g	-	-	-	g	g	-	g

Tabelle 3.7: Amplifikations- und Deletionspeaks vom neunten Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.8 Chromosom 10

Das zehnte Chromosom hat 135 534 747 bp. Das Centromer befindet sich ungefähr bei 40.2 Mbp.

Bei LAML sind wieder nur wenige Amplifikationen und Deletionen vorhanden. Auch beim THCA ist nur eine Deletion im q-Arm ab 50 Mbp und eine kleine Häufung um 90 Mbp. Dies deckt sich auch mit dem Gesamtbild des Chromosoms 10. Es ist vor allem von Deletionen betroffen, wobei in den meisten Fällen der q-Arm weitaus stärker deletiert ist. Im Bereich der Deletionsanhäufung von THCA finden wir auch einen Peak, der bei fast allen Entitäten zu finden ist (67%, vergleiche Tabelle 3.8). Besonders stark deletiert ist das zehnte Chromosom bei GBM, LGG und SKCM. Dafür finden sich bei diesen Entitäten kaum Amplifikationen. Nur bei GBM gibt es im p-Arm eine kleine Häufung von Amplifikationen. Bei OV, UCEC, BLCA, BRCA und STAD haben wir neben einer hohen Deletionsrate auch ein hohes Amplifikationsaufkommen. Dabei ist der p-Arm deutlich

mehr betroffen als der q-Arm (OV, BLCA und BRCA) beziehungsweise sind die Amplifikationen recht gleichmäßig über das Chromosom verteilt (UCEC und STAD). Neben OV, BLCA und BRCA haben auch LUSC, LUAD, KIRP und wie vorher schon erwähnt GBM einen stärker amplifizierten p-Arm. Auch bei der Amplifikation gibt es zwei auffällige Peaks. Ein Peak befindet sich im p-Arm bei 5 Mbp (61%, vergleiche Tabelle 3.8) und ein weiterer im q-Arm bei 78 Mbp (39%, vergleiche Tabelle 3.8).

	Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
5	10p15		g	g	s	s	-	-	-	-	s	g	s	s	-	s	-	s	-	s
78	10q22		s	g	-	-	-	-	-	-	-	s	-	s	-	s	-	s	-	g
89	10q23		g	g	g	g	-	-	-	s	-	s	g	g	g	-	g	g	-	g

Tabelle 3.8: Amplifikations- und Deletionspeaks vom zehnten Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

Die Muster sehen bei den doppelten Entitäten wieder sehr ähnlich aus. Nur bei OV gibt es eine Unstimmigkeit bei circa 40 Mbp. Dies wird in Abbildung 3.1 deutlich. Bei den Proben, die in das Hudson Alpha Institut geschickt wurden, gibt es einen starken Amplifikationspeak an besagter Position. In der Grafik vom Broad Institut ist dieser Peak allerdings nicht zu sehen. Die restlichen Peaks sind in diesem Fall an den gleichen Positionen. Nur die Intensität der Amplifikationen sind beim Broad Institute deutlich stärker ausgeprägt.

3.9 Chromosom 11

Das elfte Chromosom hat 135 006 516 bp. Das Centromer befindet sich ungefähr bei 53.7 Mbp.

	Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
70	11q13		g	g	s	-	g	-	s	-	-	g	g	s	s	-	s	g	-	s

Tabelle 3.9: Amplifikations- und Deletionspeaks vom elften Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

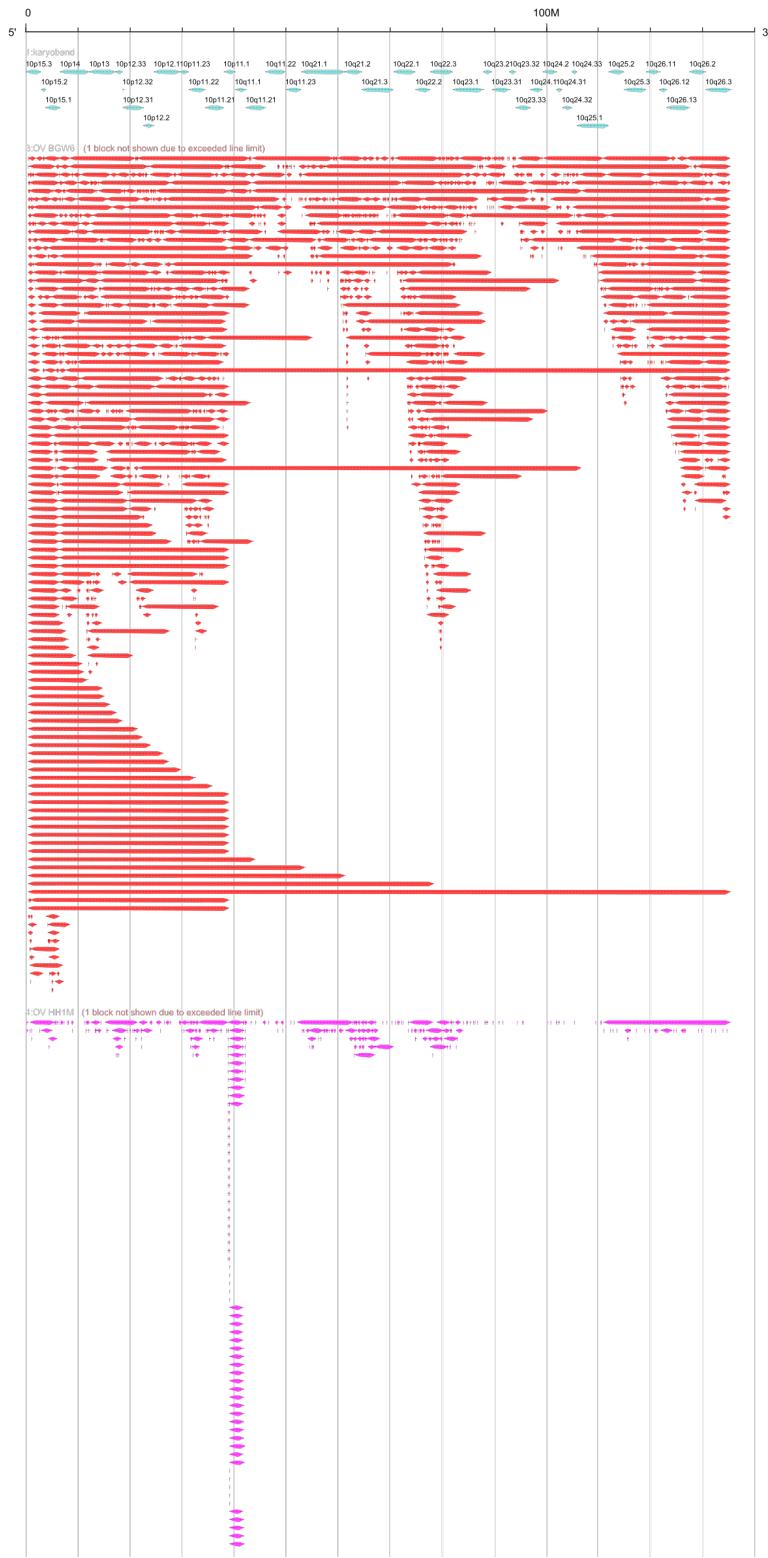


Abbildung 3.1: *FISH Oracle 2* Ausgabe von der Tumorentität OV. Track 1 ist das Chromosomenband, welches die einzelnen Chromosomabschnitte zeigt. Track 3 zeigt die Amplifikationen der OV-Tumorentität vom Broad Institute in rot. Track 4 stellt die Amplifikationen der OV-Tumorentität vom Hudson Alpha in magenta dar.

3.10 Chromosom 12

Das zwölftes Chromosom hat 133 851 895 bp. Das Centromer befindet sich circa bei 35.8 Mbp.

	Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
12	12p13		s	g	s	s	-	-	-	g	-	s	-	-	g	-	s	-	s	
26	12p11/12p12		s	s	-	s	s	s	-	-	s	g	s	g	-	s	s	g	-	g
70	12q15		g	g	-	g	s	-	-	-	s	g	g	g	-	-	g	g	-	g

Tabelle 3.10: Amplifikations- und Deletionspeaks vom zwölften Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.11 Chromosom 16

Das 16. Chromosom hat 90 354 753 bp. Das Centromer befindet sich circa bei 36.6 Mbp.

	Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
79	16q23		-	g	g	-	g	-	-	s	-	-	g	-	-	g	-	s	g	

Tabelle 3.11: Deletionspeaks vom 16. Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.12 Chromosom 17

Das 17. Chromosom hat 81 195 210 bp. Das Centromer befindet sich circa bei 24 Mbp.

3.13 Chromosom 19

Das 19. Chromosom hat 59 128 983 bp. Das Centromer befindet sich ungefähr bei 26.5 Mbp.

	Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
12	17q23		s	g	s	-	-	-	-	-	-	-	s	-	-	-	g	-	-	
38	17q12/17q21		s	g	s	s	g	-	-	-	-	-	g	s	-	s	-	g	-	g

Tabelle 3.12: Deletionspeaks vom 17. Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

Bei den doppelten Entitäten von Hudson Alpha, OV und GBM, kommt es im Bereich um das Centromer zu einem Amplifikationsblock. Dieser beinhaltet in beiden Fällen die kompletten Abschnitte 19p11, 19q11 und 19q12 und ist sehr stark ausgeprägt. Dies tritt nur bei den Proben von Hudson Alpha auf und wird in der Abbildung 3.2 veranschaulicht.

	Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
30.5	19q12		g	g	s	s	-	-	-	-	-	-	g	s	g	-	-	g	-	g

Tabelle 3.13: Amplifikationspeaks vom 19. Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.14 Chromosom 20

Das 20. Chromosom hat 63 025 520 bp. Das Centromer befindet sich ungefähr bei 27.5 Mbp.

	Position (Mbp)	Chromosom-abschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
15	20p12		s	g	g	-	s	s	-	-	-	-	g	s	s	g	-	g	g	
30	20q11		g	s	s	-	s	-	-	-	s	s	g	g	s	s	-	g	g	

Tabelle 3.14: Amplifikations- und Deletionspeaks vom 20. Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.



Abbildung 3.2: *FISH Oracle* Ausgabe vom Chromosom 19. Dargestellt wird der Bereich zwischen 20 Mbp und 40 Mbp. Track 3 zeigt die Amplifikationen vom GBM und Track 4 von OV. In beiden Fällen stammen die Daten von Hudson Alpha, aber aus unterschiedlichen *SNP-Microarrays*. In beiden Fällen ist um das Centromer (26.5) ein Amplifikationsblock zu erkennen.

3.15 Chromosom X

Das X-Chromosom hat 155 270 560 bp. Das Centromer befindet sich ungefähr bei 60.6 Mbp.

	Position (Mbp)	Chromosomabschnitt	BLCA	BRCA	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC	
11	Xp22		s	-	-	s	s	-	-	-	-	s	s	s	g	-	-	s	g	-	-
32	Xp21		-	-	-	g	-	-	-	-	-	s	s	g	s	-	s	g	-	g	

Tabelle 3.15: Deletionspeaks vom X-Chromosom. Der größte Peak einer Entität ist **fett** gesetzt.

3.16 Chromosom Y

Die Ergebnisse von der LAML Entität, welche mit dem *SNP-Microarray WGW6* gemessen wurden, enthalten als einzige Daten für das Y-Chromosom. In diesem Fall sind es nur Deletionsdaten. Bei den anderen Entitäten wird das Y-Chromosom nicht verändert.

3.17 Analyse der häufigsten Peaks

Zur weiteren Analyse werden die Chromosomabschnitte der sieben häufigsten Peaks genauer untersucht. Diese sind der Tabelle 3.16 zu entnehmen. Von besonderem Interesse sind hierbei die Gene, die in dem Bereich der Deletion beziehungsweise Amplifikation liegen.

Zu dieser Analyse wurde wieder die Software *FISH Oracle 2* verwendet. Zunächst wurde nach dem betreffendem Abschnitt im entsprechendem Chromosom gesucht. Dabei wurde die Region von 2 Mbp vor bis 2 Mbp hinter der Position des zu betrachtenden Peaks mit *FISH Oracle 2* visualisiert. In dieser Auflösung werden zusätzlich auch die Gene dieses Abschnittes in *FISH Oracle 2* mit dargestellt. Für die Suche wurden alle Entitäten, die den jeweiligen Peak enthalten, in einem Track zusammengefasst. Dabei entstanden bei sechs der so näher betrachteten Chromosomabschnitte große Peaks, die auf Genen liegen, die für die Krebsforschung relevant sind.

Die Gene, die in dem Bereich der großen Peaks liegen, sind ebenfalls in der Tabelle 3.16 dargestellt. Diese Gene wurden anhand von [20] und [13] auf ihre Relevanz in der Tumorforschung überprüft (siehe Kapitel 4). Diese Gene liegen vermutlich der Deletion

Häufigkeit	Abschnitt	GOI	Entitäten
72%	9p21	MTAP CDKN2A CDKN2B	BLCA, BRCA, COAD, GBM, HNSC, KIRC, LGG, LUAD, LUSC, OV, SKCM, STAD, UCEC
72%	12p11 + 12p12	KRAS	BLCA, BRCA, GBM, HNSC, KIRC, LGG, LUAD, LUSC, OV, READ, SKCM, STAD, UCEC
72%	20q11	BCL2L1	BLCA, BRCA, COAD, HNSC, LGG, LUAD, LUSC, OV, PRAD, READ, SKCM, STAD, UCEC
67%	2q22	LRP1B	BLCA, BRCA, COAD, GBM, HNSC, KIRC, LUAD, LUSC OV, PRAD, SKCM, UCEC
67%	10q23	PTEN	BLCA, BRCA, COAD, GBM, LAML, LUAD, LUSC, OV, PRAD, SKCM, STAD, UCEC
67 %	11q13	ORAOV1 FGF19	BLCA, BRCA, COAD, HNSC, KIRP, LUAD, LUSC, OV, PRAD, SKCM, STAD, UCEC
67%	20p12		BLCA, BRCA, COAD, HNSC, KIRC, LUSC, OV, PRAD, READ, STAD, THCA, UCEC

Tabelle 3.16: Die sieben am häufigsten in den Chromosomen auftretenden Peaks. Sofern der Abschnitt eines oder mehrere Gene von Interesse (*Genes of Interest, GOI*) für die Krebsforschung beinhaltet, werden diese zu den Peaks aufgelistet. Die Häufigkeit gibt an, bei wie vielen Entitäten der Peak auftritt. Die Farbe zeigt an, ob es sich um eine Deletion (blau) oder Amplifikation (rot) handelt.

beziehungsweise Amplifikation in diesen Bereichen zugrunde.

Bei dem Chromosomabschnitt 20q11 lagen die zwei entstandenen Peaks zwar in dem Bereich der *Macro Domain* MACROD2, allerdings wurde die *Macro Domain* nicht komplett von den Peaks abgedeckt.

4 Diskussion

Im Verlauf der Arbeit haben sich einige Probleme ergeben. Unter anderem ist es nur schwierig möglich alle Grafiken aller Tumorentitäten für ein Chromosom übersichtlich darzustellen. Dies liegt zum einen an den unterschiedlichen Höhen der Grafiken, da sich diese nach den größten Peaks richtet. Bei den drei Tumorentitäten, die in unterschiedlichen Instituten getestet wurden (GBM, LAML und OV) ist die Intensität der CNVs bei den Ergebnissen des Broad Institutes deutlich höher. Dieses Problem hätte mit einem individuellen *intensity*-Wert bei der Visualisierung mit *FISH Oracle 2* sehr einfach lösen können. Deutlich wird dies bei einem Vergleich von Abbildung 3.1 und Abbildung 4.1. In Abbildung 4.1, sieht das Amplifikationsmuster zwischen beiden Instituten schon deutlich ähnlicher aus. Dies liegt an einem niedrigeren *intensity*-Wert bei den Daten vom Hudson Alpha. Aufgrund der großen Anzahl an zu erzeugenden Bildern (485 Stück) wurde sich aber für einen globalen *intensity*-Wert von +/- 0.5 entschieden. Der einheitliche *intensity*-Wert ist auch der Grund, weshalb bei einigen Tumorentitäten kaum CNVs zu erkennen sind. Dies ist vor allem bei LAML, THCA, KIRC und KIRP der Fall.

Außerdem gibt es bei den Daten der Hudson Alpha immer wieder ganze Blöcke (siehe Kapitel 3.13), die im Vergleich dazu bei den Daten des Broad Institute nicht zu sehen sind. Eine Ursache hierfür könnte die Wahl der Dateien gewesen sein. Von den Hudson Alpha Proben wurde die Dateiendung *.seg.txt* zur Weiterverarbeitung verwendet (siehe Auflistung 2.1).

In Kapitel 3 wird deutlich, dass sich die Entitäten anhand ihrer Muster und Peaks nicht in eindeutige Gruppen aufteilen lassen. Eine Klassifizierung pro Chromosomen zu erstellen, war ebenfalls nicht möglich. Der Versuch, wie bei Chromosom 4 und 5 (siehe Kapitel 3.3 und Kapitel 3.4) zusätzlich noch zwischen Amplifikationen und Deletionen bei der Einteilung in Gruppen zu unterscheiden, klappt nur bedingt. Denn einige Entitäten würden sich auch mehreren Gruppen zuordnen lassen. Entitäten einer bestimmten Gruppe können auch die Merkmale einer anderen Gruppe enthalten. Somit gibt es immer wieder Entitäten, welche sich zwei oder mehr Gruppen zuordnen lassen können. Zwischen den Gruppen lassen sich also keine eindeutigen Grenzen ziehen. So wurden die beiden Tumorentitäten BRCA und HNSC für ihre Deletionsmuster von Chromosom 4 in zwei unterschiedliche Gruppen aufgeteilt. Betrachtet man nur die zwei Grafiken der beiden Entitäten miteinander, so weisen sie ein sehr ähnliches Deletionsmuster auf (siehe Abbildung 4.2 und Abbildung 4.3). Bei beiden Entitäten nimmt die Deletionsdichte von dem Armenden zum Centromer hin ab. Dieses Beispiel zeigt, das dadurch eine Klassifizierung erheblich erschwert wird und es wurde im weiteren Verlauf der Arbeit davon abgesehen, alle Chromosomen zu gruppieren.

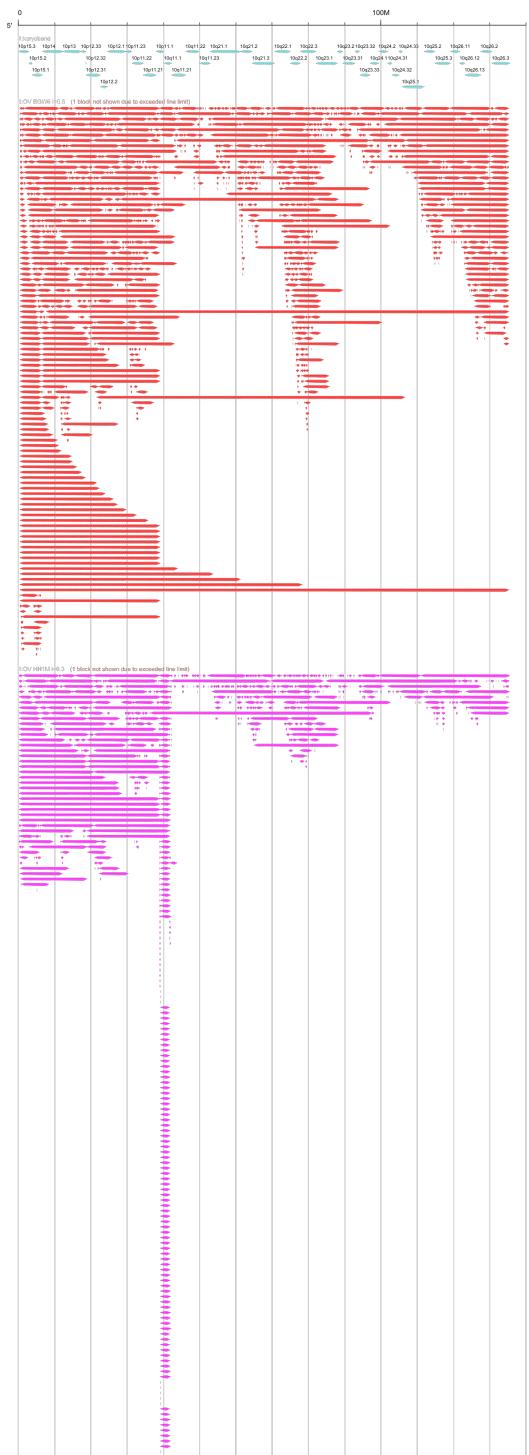


Abbildung 4.1: *SVG FISH Oracle* Ausgabe: Amplifikationen von Eierstockkrebspatientinnen (OV) vom BGW6 und HH1M auf Chromosom 10. Track 1 ist das Chromosomenband (karyoband), welches die einzelnen Chromosomabschnitte zeigt. Track 3 zeigt die Amplifikationen der OV-Tumorentität vom Broad Institute in rot mit einem *intensity*-Wert von 0.5. Track 4 stellt die Amplifikationen der OV-Tumorentität vom Hudson Alpha in Magenta mit einem *intensity*-Wert von 0.3 dar.

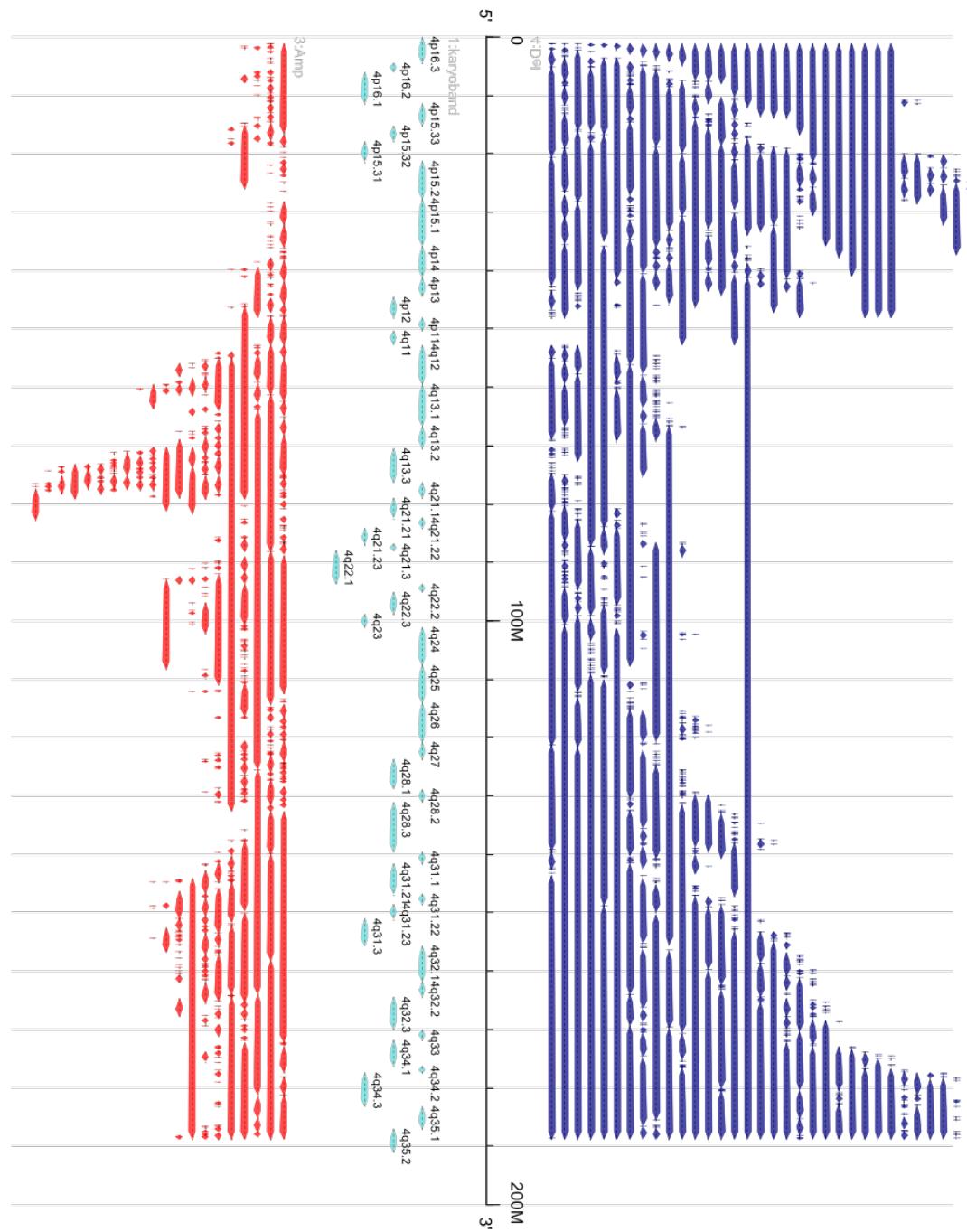


Abbildung 4.2: *SVG FISH Oracle* Ausgabe: CNVs von Brustkrebspatientinnen (BRCA) auf Chromosom 4. Track 1 ist das Chromosomenband (karyoband), welches die einzelnen Chromosomabschnitte zeigt. Track 3 zeigt die Amplifikationen in rot. Track 4 stellt die Deletionen in blau dar.

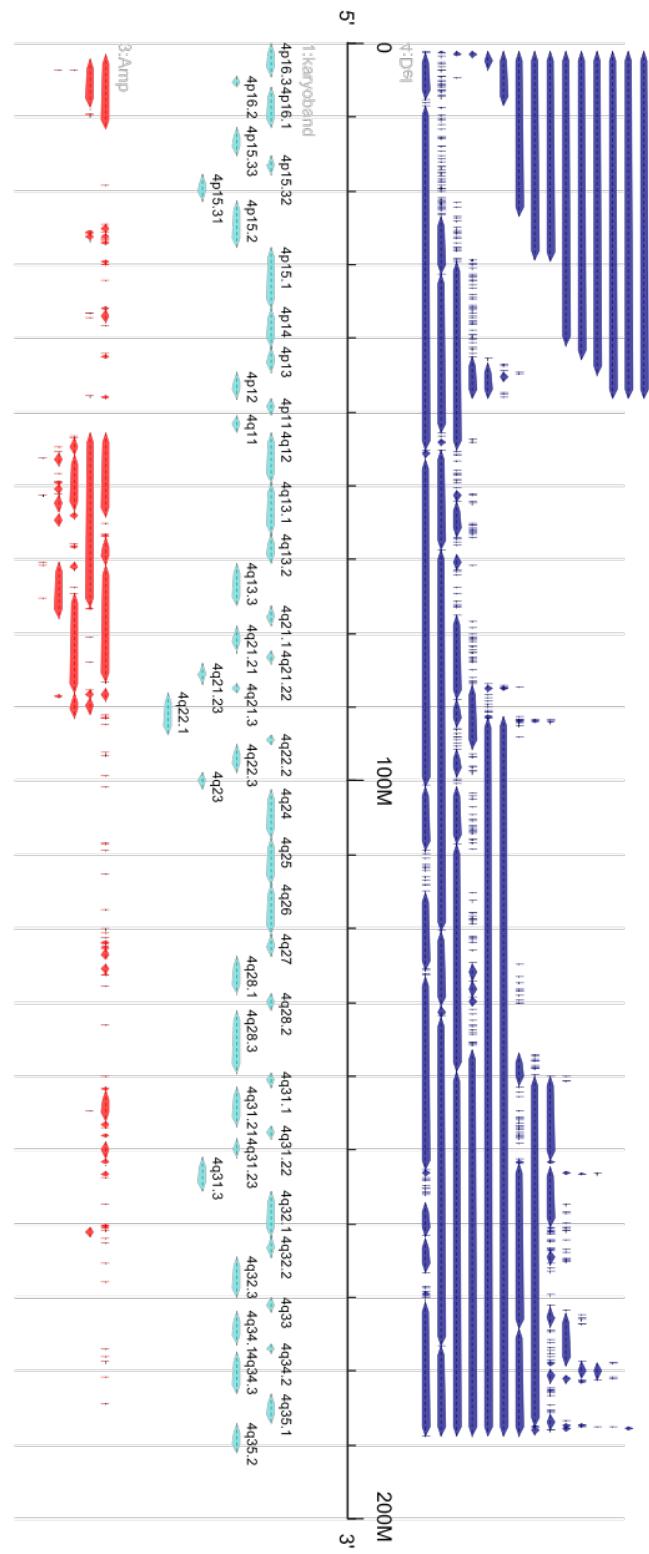


Abbildung 4.3: *SVG FISH Oracle* Ausgabe: CNVs von Patienten mit Kopf-Hals-Karzinom (HNSC) auf Chromosom 4. Track 1 ist das Chromosomenband (karyoband), welches die einzelnen Chromosomabschnitte zeigt. Track 3 zeigt die Amplifikationen in rot. Track 4 stellt die Deletionen in blau dar.

Allerdings wurde eine gute Übersicht erstellt, womit gezielt bestimmte Bereiche von zwei oder mehreren Entitäten verglichen werden können. Deshalb sind die auffälligen Peaks mehr in den Fokus gerückt und wurden tabellarisch, wenn vorhanden, für jedes Chromosom aufgelistet (siehe Tabelle 3.16). Diese wie in Kapitel 3.17 gefundenen Gene sind alle relevant für die Krebsforschung[13][20]. Um dies zu zeigen, gehe ich im Folgenden auf die Gene der ersten drei Abschnitte genauer ein.

Im Abschnitt 9p21 wurden die Gene CDKN2A und CDKN2B identifiziert. Diese zwei Gene codieren die Proteine p16 und p15, CDK-Inhibitoren (aus dem Englischen *cyclin-dependent kinase*), welche durch eine starke Bindung an CDK4 und CDK6 den Zellzyklus regulieren. Durch die Deletion in diesem Abschnitt, sind die exprimierten CDK-Inhibitoren häufig ineffizient und somit die Regulierung des Zellzyklus nicht mehr gewährleistet. Die räumliche Nähe von MTAP zu diesen Genen, führt häufig zu einer Co-deletion. Die 5-Methylthioadenosin-Phosphorylase (MTAP) wird im Methionin-Salvage-Stoffwechsel benötigt und katalysiert dabei die Phosphorylierung von 5-Methylthioadenosin (MTA) unter Abspaltung von Adenosin.

Das KRAS-Gen wurde im Abschnitt 12p11 und 12p12 identifiziert. Dieses Onkogen spielt ebenfalls eine wichtige Rolle beim Zellzyklus und gehört zur Familie der GTPasen. Es ist also ein GTP-bindendes Protein, ein sogenanntes G-Protein. Die Amplifikation dieses Gens führt sehr wahrscheinlich zu einer Überexpression und somit zu einem Ungleichgewicht beim Stoffwechsel der Zelle, da dieses Protein an vielen Signaltransduktionen beteiligt ist.

Auch das im Abschnitt 20q11 gefundene Gen BCL2L1 spielt eine große Rolle in der Krebsforschung. Es codiert das Protein Bax, welches als Co-Faktor des Tumorsuppressor-Proteins p53 wirkt. In Verbindung mit p53 beschleunigt es die Einleitung zur Apoptose. Auch hier löst die Vervielfältigung des Gens wahrscheinlich eine Überexpression aus.

In dem letzten Abschnitt (20p12) konnte kein Gen identifiziert werden, da es keinen eindeutigen Peak gab. Dies liegt zu einer hohen Wahrscheinlichkeit an der Makrodomäne MACROD2.

5 Ausblick

Die in der vorliegenden Arbeit verfolgte Hypothese, Tumorentitäten anhand von Amplifikations- und Deletionsmustern zu gruppieren erscheint vielversprechend. Um größere Datenmengen verarbeiten zu können ist eine Automatisierung des Vorgangs notwendig. Die Beschreibung des Programms *GISTIC*[11] legt einen hohen Grad an Automatisierung nahe, was auch in diesem Ansatz wünschenswert wäre.

Hilfreich bei der Arbeit mit *FISH Oracle 2* wäre entweder ein API oder eine Art Batch-Betrieb, der die Weiterverarbeitung der gewonnenen Diagramme erleichtert.

Die im Rahmen dieser Arbeit notwendige Durchsicht der Daten unterstreicht, dass die Aufbereitung und Visualisierung von derartigen Datenmengen eine große Herausforderung ist. Die Entwicklung von Software zur interaktiven Analyse von Genomsequenzdaten wie *FISH Oracle 2* wird eine wachsende Rolle in der Bioinformatik und den Lebenswissenschaften spielen.

Die CNV-Daten sind nur ein kleiner Teil der großen *TCGA*-Datensammlung. Auch die Visualisierung dieser Daten in *FISH Oracle 2* ist nur ein Teil des Funktionsumfangs. Es könnten beispielsweise noch Tranlokationsdaten oder *Single-Nucleotide-Variations (SNV)* eingesetzt werden, um das Bild über die Tumorentitäten abzurunden. Zusätzlich bietet der *TCGA* auch noch klinische Daten an, die für eine tiefere Analyse interessant werden könnten.

Die Verwendung von Daten aus verschiedenen Quellen erwies sich als problematisch. Eine stärkere Normalisierung der Daten erscheint notwendig, beispielsweise durch eine individuelle Anpassung des *intensity*-Wertes pro Institut oder *SNP-Microarray*. Außerdem sollte getestet werden ob mit den normalisierten CNV-Daten (*.segnormal.txt*) der Hudson Alpha immer noch Blockamplifikationen, wie in Abbildung 3.2 in Kapitel 3.13 beschrieben auftreten.

Bei der Untersuchung der Daten auf verschiedenen Skalen könnten schwächere Peaks erkannt werden, die bei der in dieser Arbeit gewählten Auflösung nicht sichtbar wurden.

A Anhang

Es folgen eine Aufschlüsselung der Kürzel der untersuchten Tumorentitäten sowie die generierten Diagramme für die drei am häufigsten auftretenden Peaks nach Tabelle 3.16 getrennt nach den einzelnen Tumorentitäten in Chromosom neun, zwölf und 20.

Alle Diagramme liegen in digitaler Form auf der CD bei und sind unter <https://github.com/bazty/ba> zu beziehen.

Kürzel	englisch	Tumorentität	importierte Datensätze
	deutsch		
READ	Rectum Adenocarcinoma	Darmkrebs	165
THCA	Thyroid Carcinoma	Schilddrüsencarzinom	497
BLCA	Bladder Urothelial Carcinoma	Harnblasenkrebs	250
COAD	Colon Adenocarcinoma	Darmkrebs	461
LUSC	Lung Squamous Cell Carcinoma	Lungenkrebs	505 + 22 (HH1M)
OV	Ovarian Serous Cystadenocarcinoma	Eierstockkrebs	572 + 516 (HH1M)
PRAD	Prostate adenocarcinoma	Prostatakrebs	373
ESCA	Esophageal carcinoma	Speiseröhrenkrebs	126
HNSC	Head and neck squamous cell carcinoma	Kopf-Hals-Karzinom	511
KIRC	Kidney renal clear cell carcinoma	Nierenkrebs	556
SKCM	Skin Cutaneous Melanoma	Hautkrebs	384
SARC	Sarcoma	Sarkom (Binde-, Stütz- & Muskelgewebe)	169
UCS	Uterine Carcinosarcoma	Müllerscher Mischtumor	56
STAD	Stomach adenocarcinoma	Magenkrebs	369
KICH	Kidney Chromophobe	Nierenkrebs	66
KIRP	Kidney renal papillary cell carcinoma	Nierenkrebs	212
LGG	Brain Lower Grade Glioma	Gliom (Hirntumor)	486
UCEC	Uterine Corpus Endometrial Carcinoma	Gebärmutterhalskrebs	524
PAAD	Pancreatic adenocarcinoma	Pankreaskrebs	102
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	Plattenepithel- und Gebärmutterhalskrebs	203
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	Lymphdrüsencarzinom	28
GBM	Glioblastoma multiforme	Glioblastom (Hirntumor)	574 + 427 (HHH5)
LUAD	Lung adenocarcinoma	Lungenkrebs	504
MESO	Mesothelioma	Mesotheliom (unlokaliert)	37
PCPG	Pheochromocytoma and Paraganglioma	Nebennierenkrebs	159
BRCA	Breast invasive carcinoma	Brustkrebs	1043
ACC	Adrenocortical carcinoma	Nebennierenrindenkarzinom	90
LAML	Acute Myeloid Leukemia	Leukämie	191 + 200 (WGW6)

Tabelle A.1: Quelldatensätze mit Kürzel, englischer und deutscher Bezeichnung und Anzahl. Nicht verwendete Datensätze sind rot markiert.

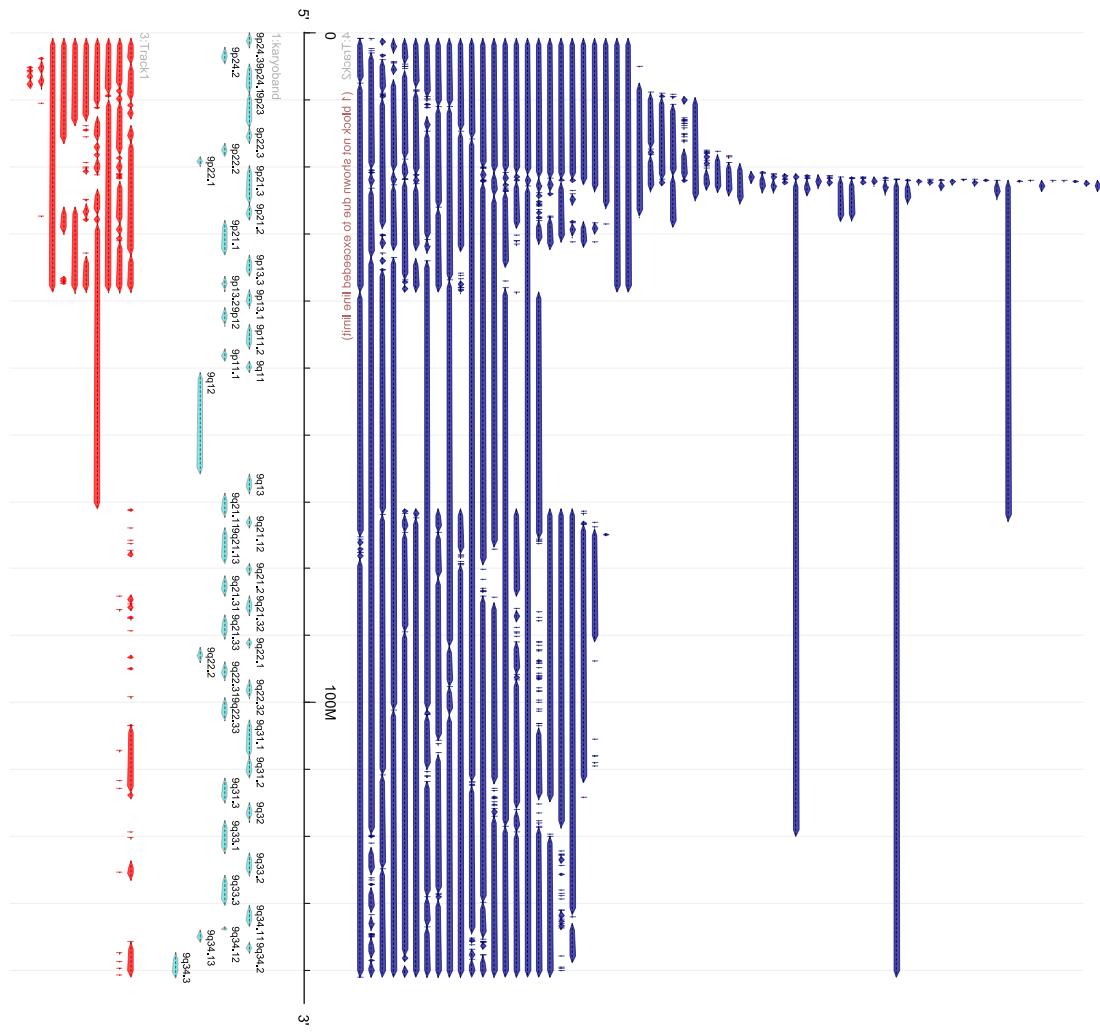


Abbildung A.1: Chromosom 9, BLCA, BGW6

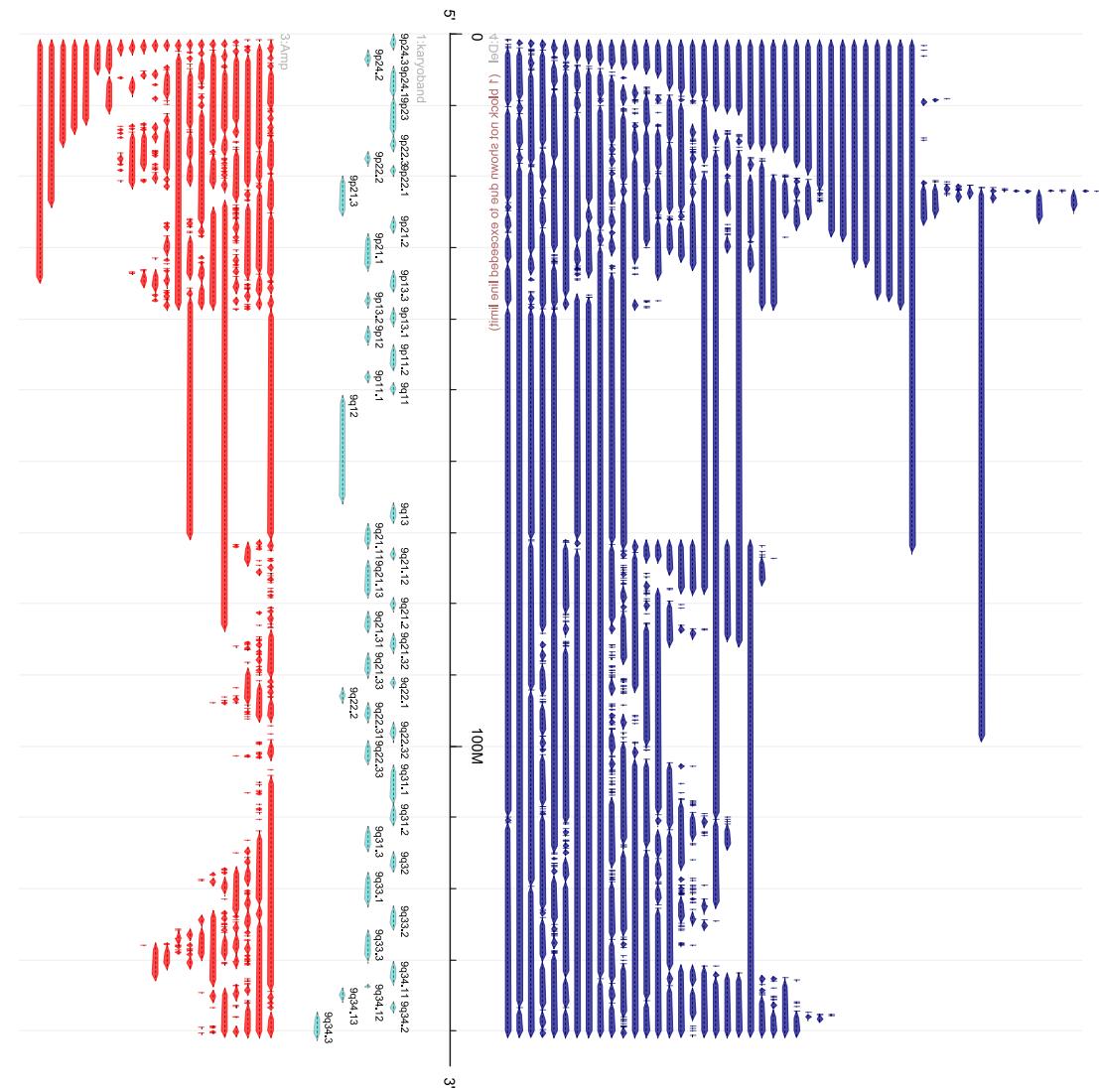
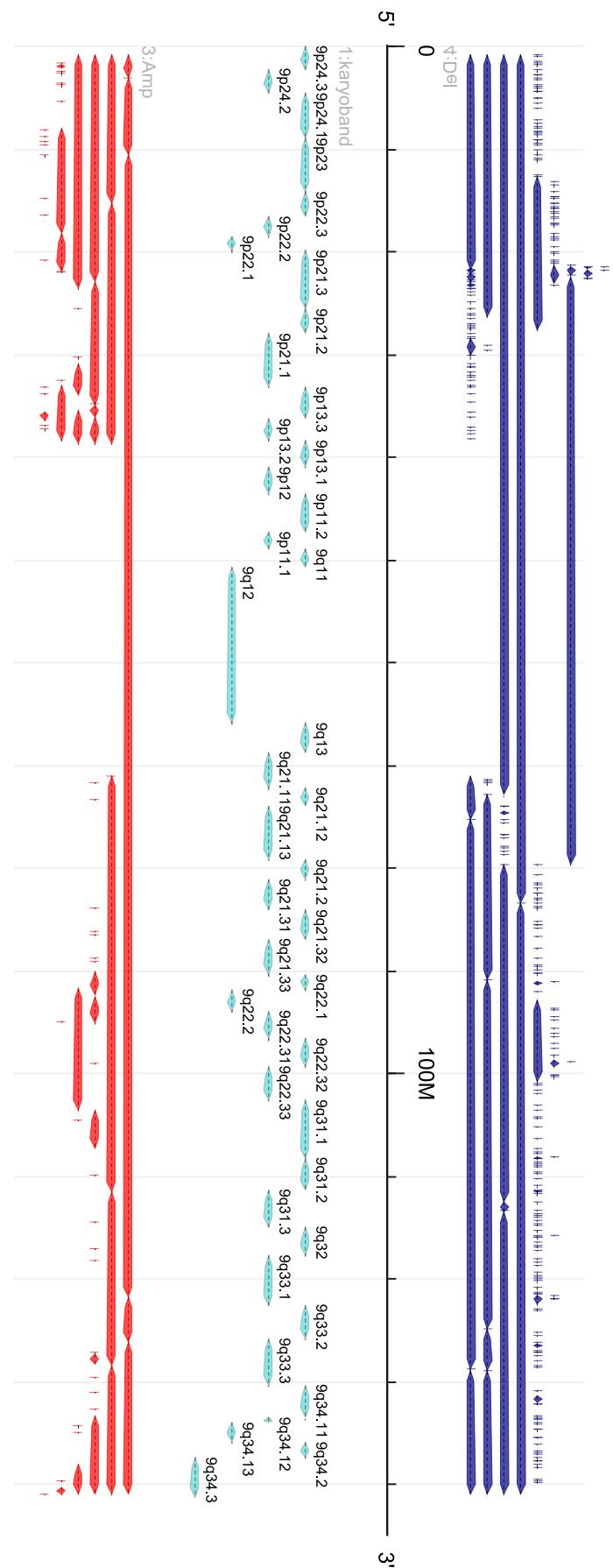


Abbildung A.2: Chromosom 9, BRCA, BGW6



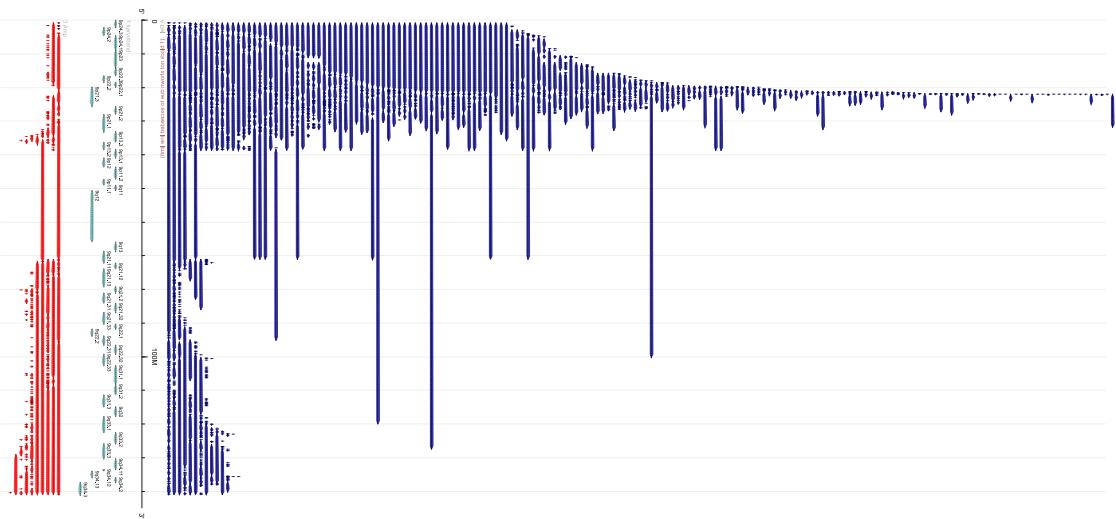


Abbildung A.4: Chromosom 9, GBM, BGW6

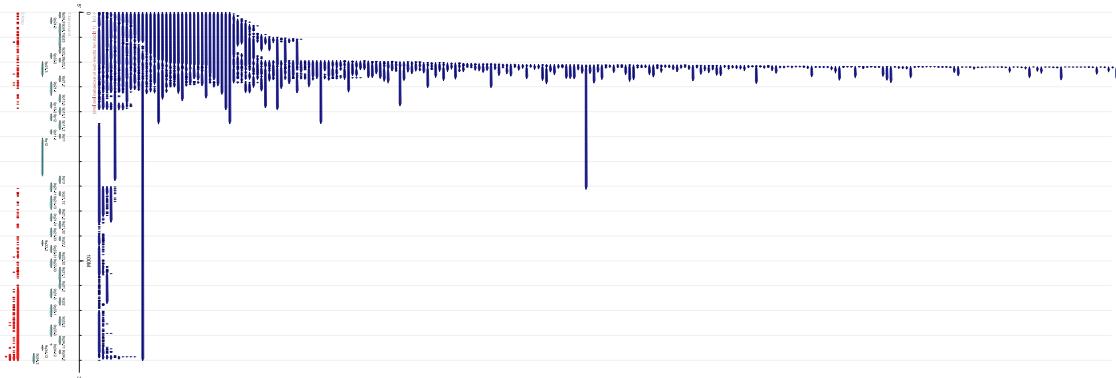


Abbildung A.5: Chromosom 9, GBM, HHH5

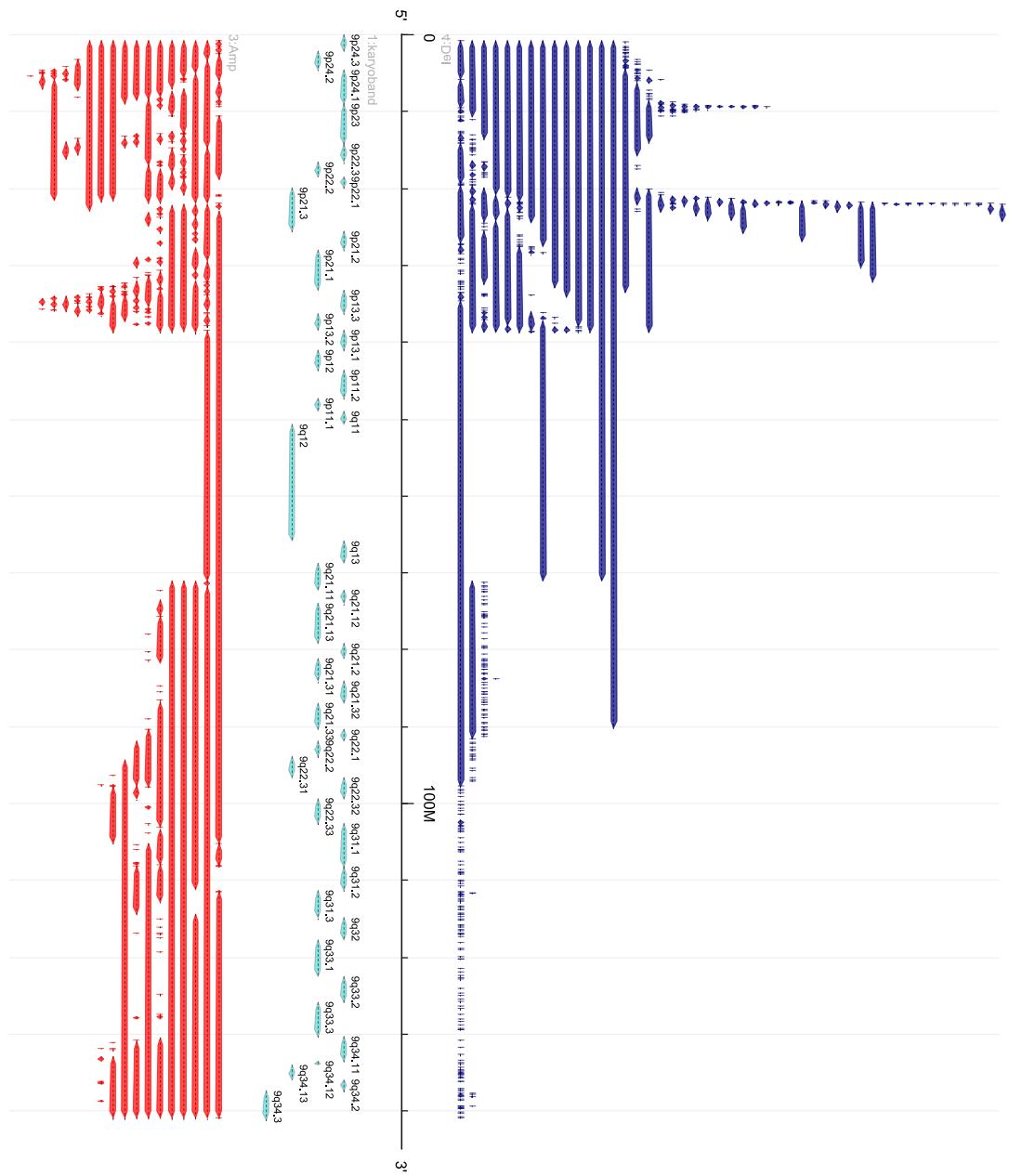


Abbildung A.6: Chromosom 9, HNSC, BGW6

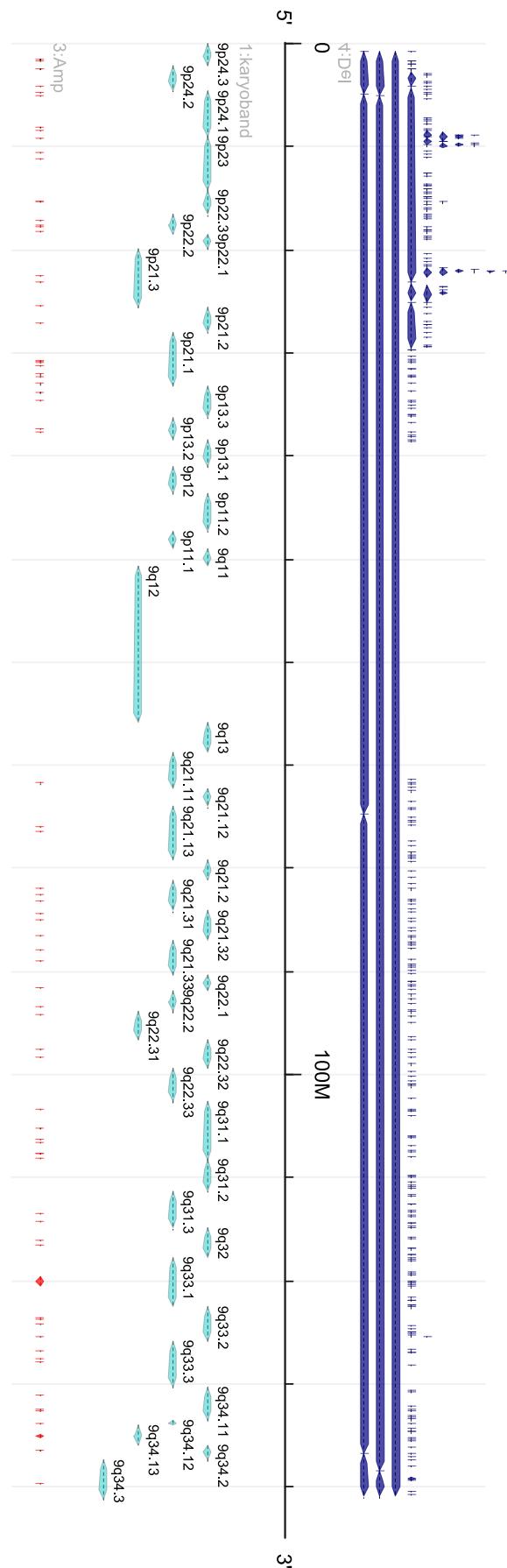


Abbildung A.7: Chromosom 9, KIRC, BGW6

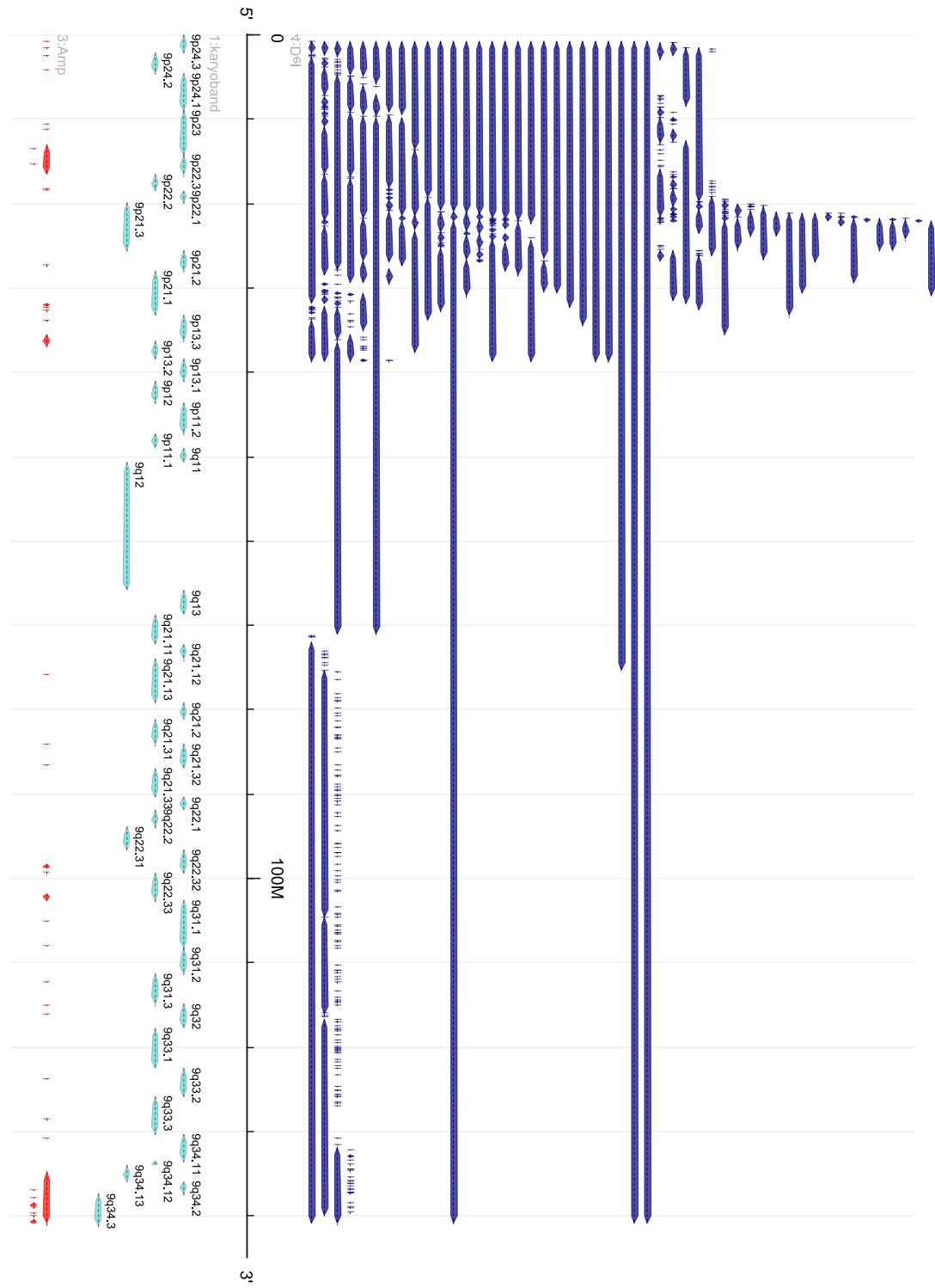


Abbildung A.8: Chromosom 9, LGG, BGW6

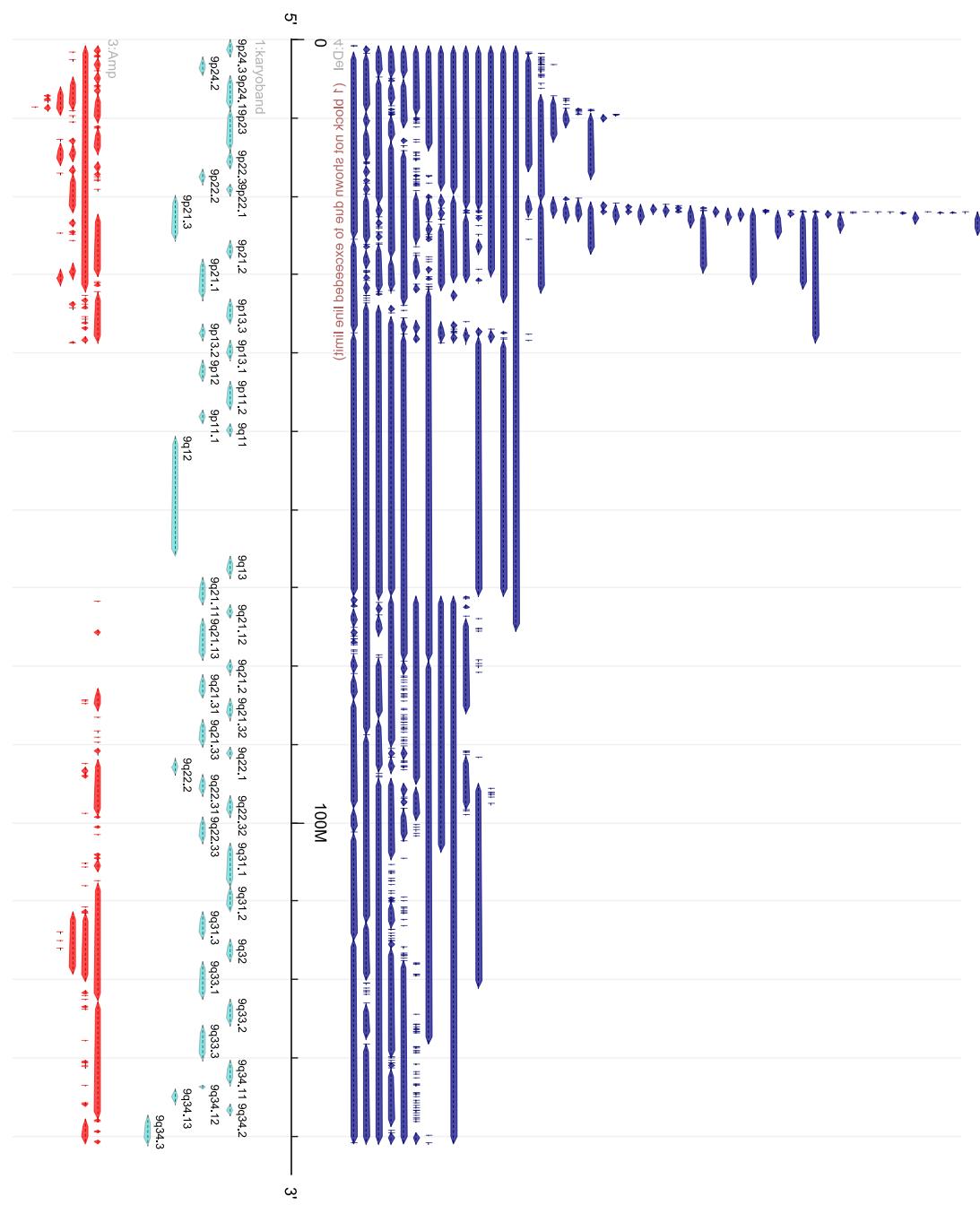


Abbildung A.9: Chromosom 9, LUAD, BGW6

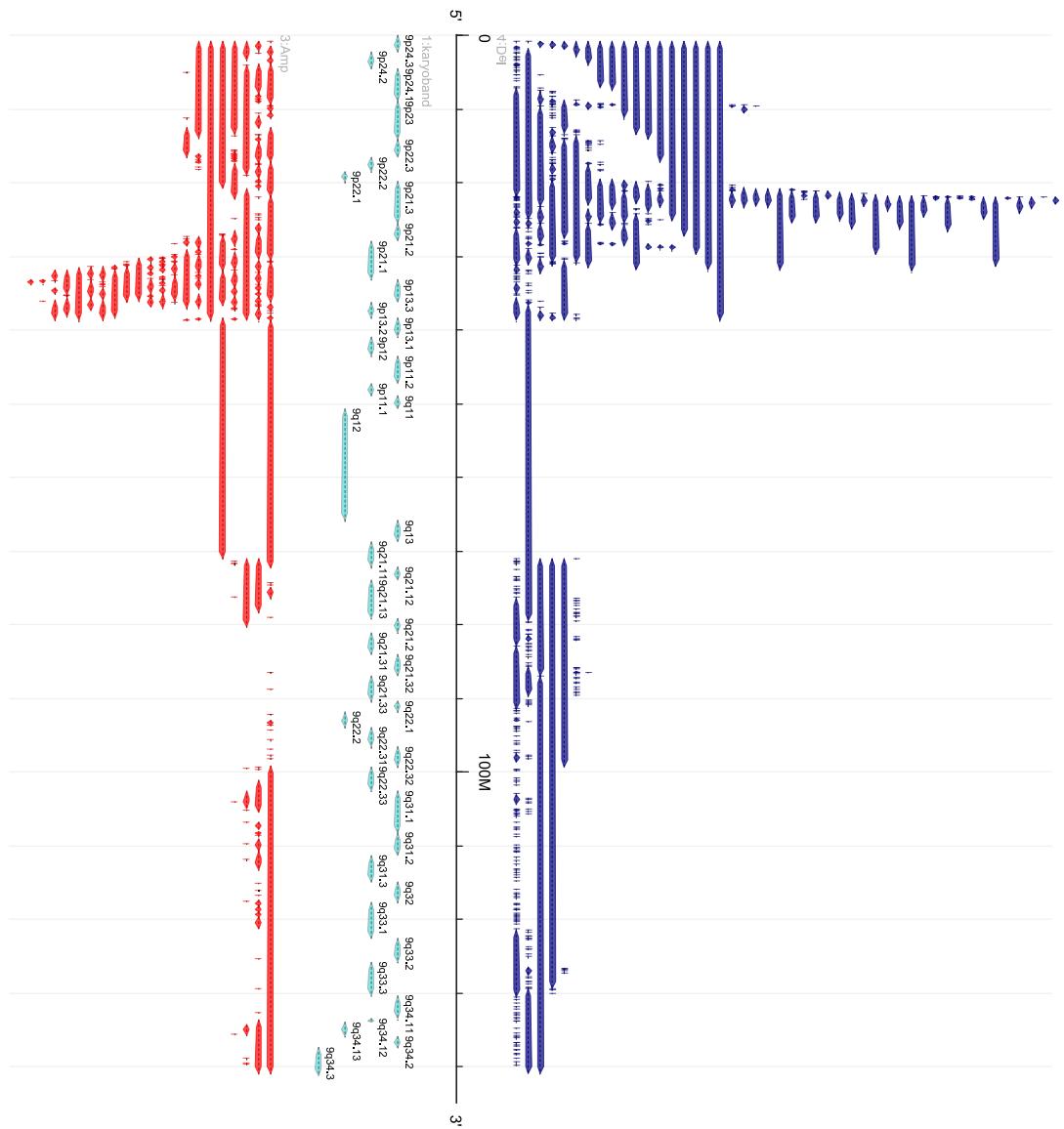


Abbildung A.10: Chromosom 9, LUSC, BGW6

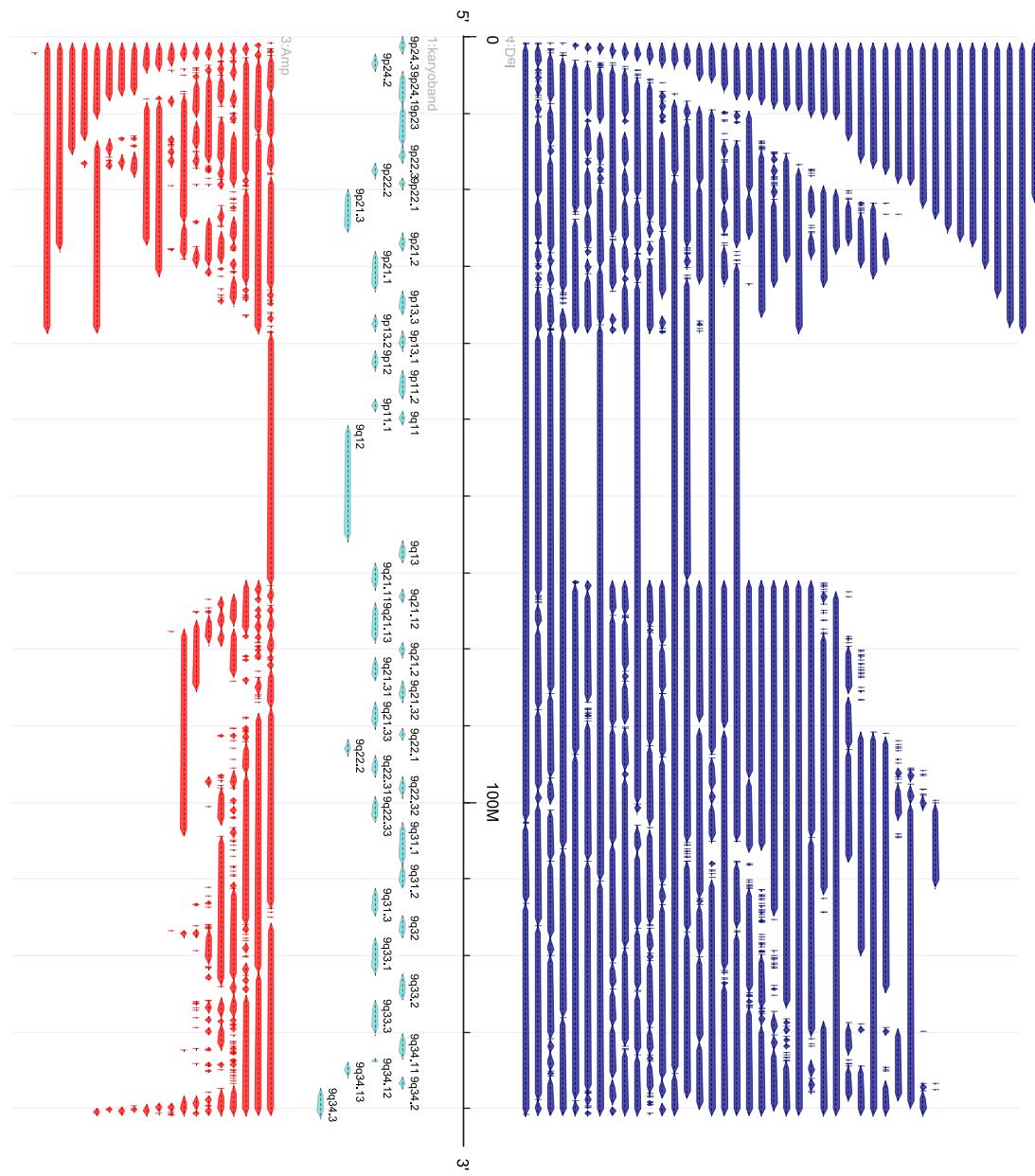


Abbildung A.11: Chromosom 9, OV, BGW6

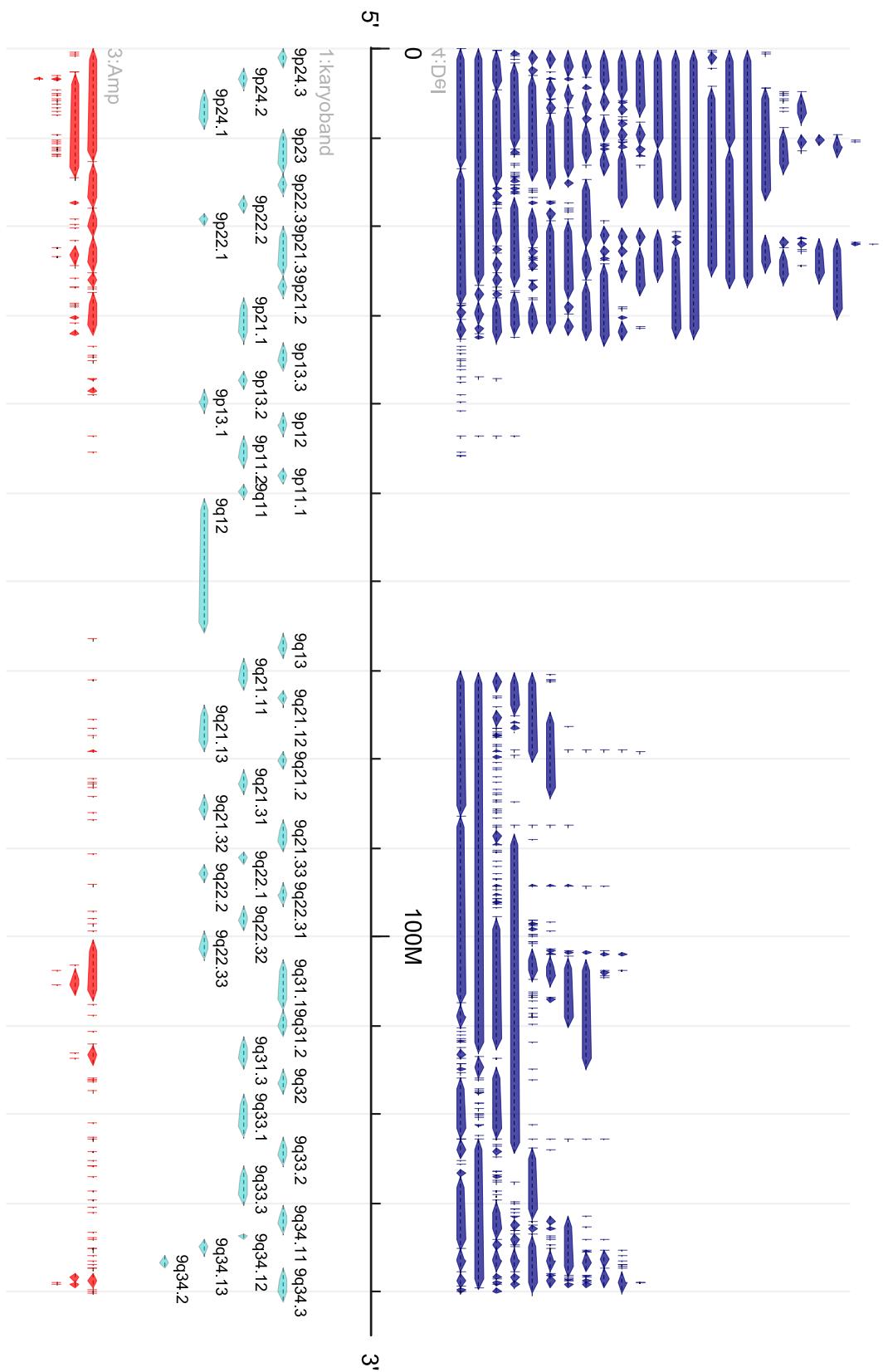


Abbildung A.12: Chromosom 9, OV, HH1M

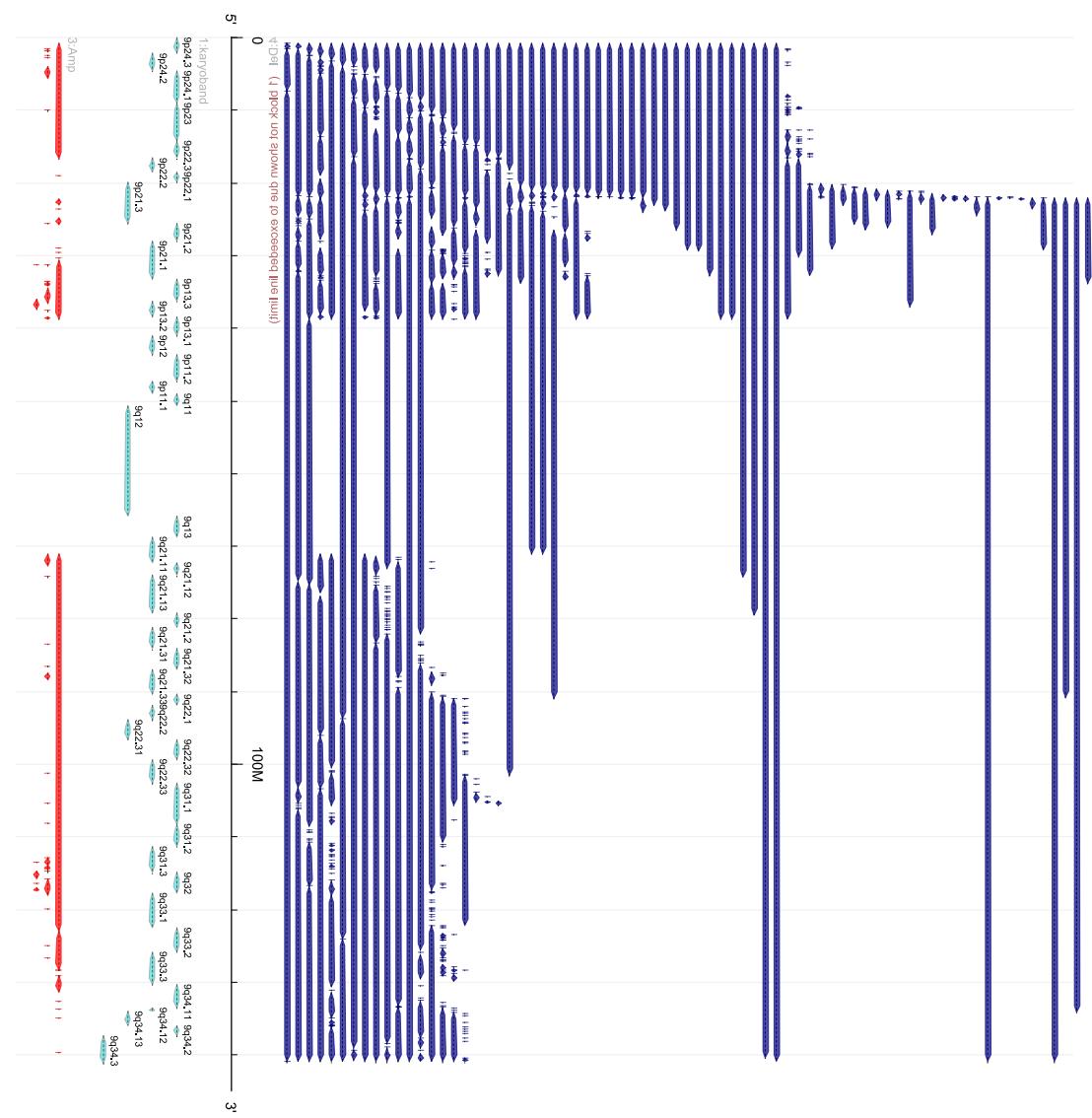


Abbildung A.13: Chromosom 9, SKCM, BGW6

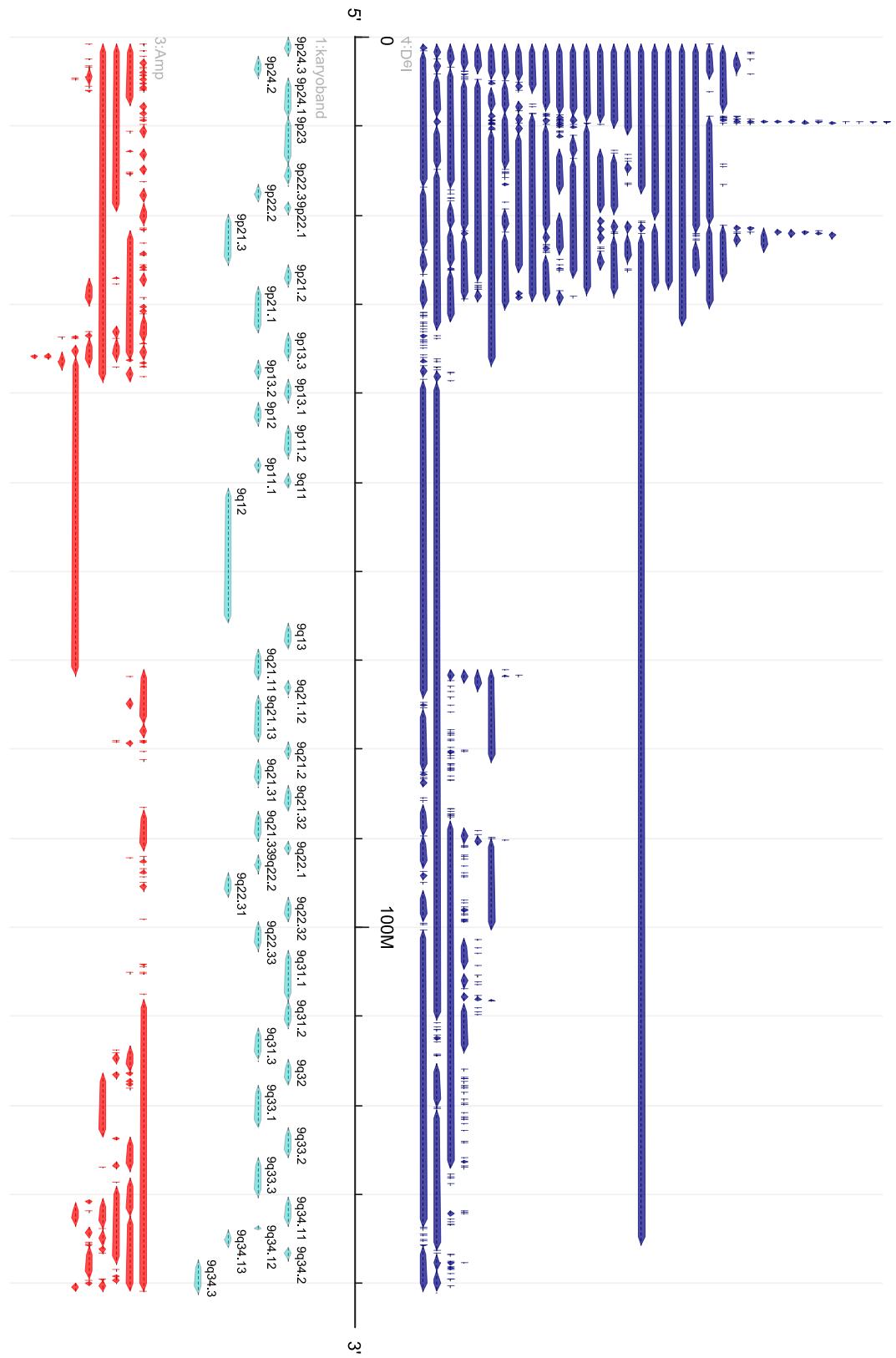


Abbildung A.14: Chromosom 9, STAD, BGW6

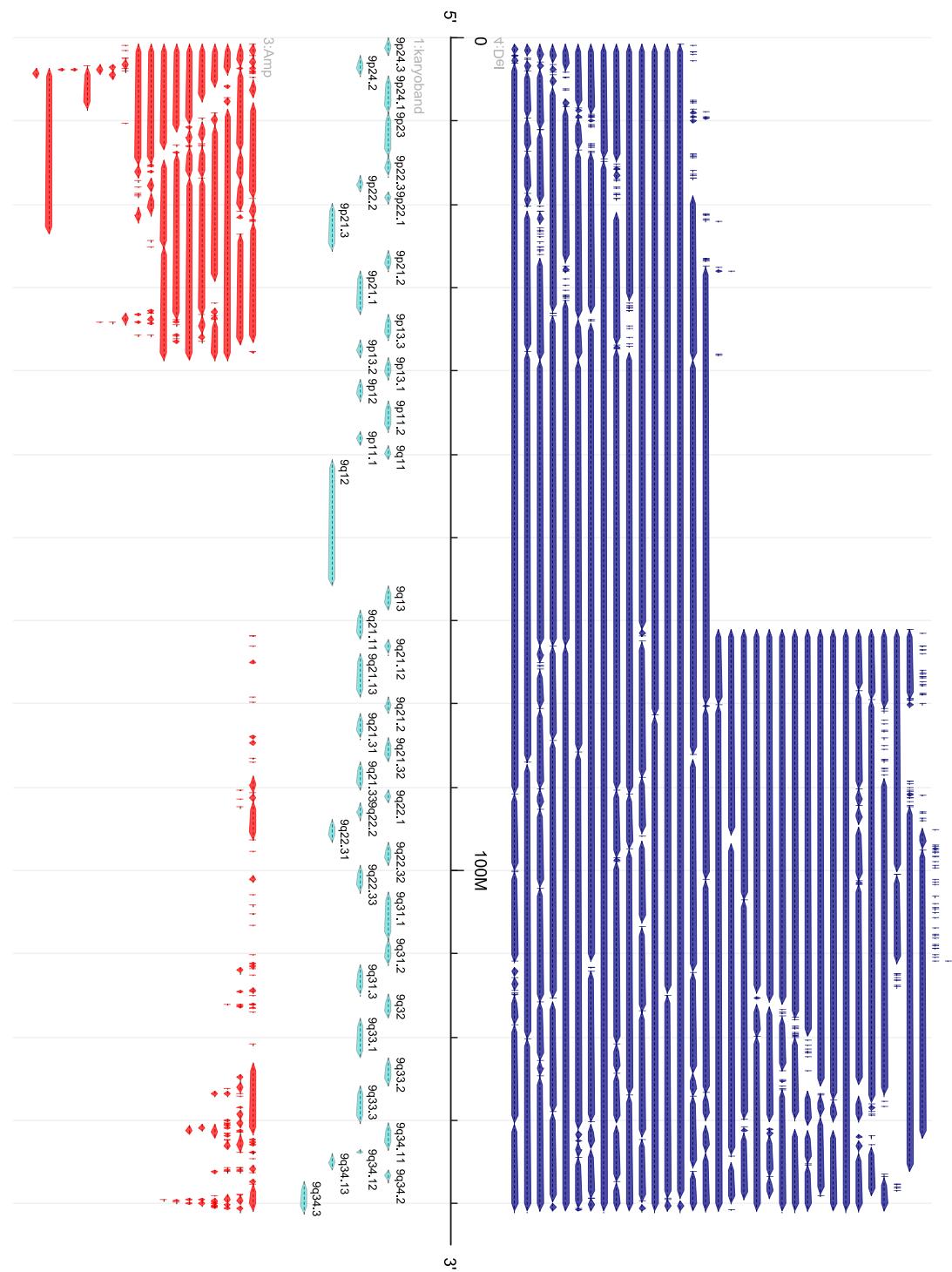


Abbildung A.15: Chromosom 9, UCEC, BGW6

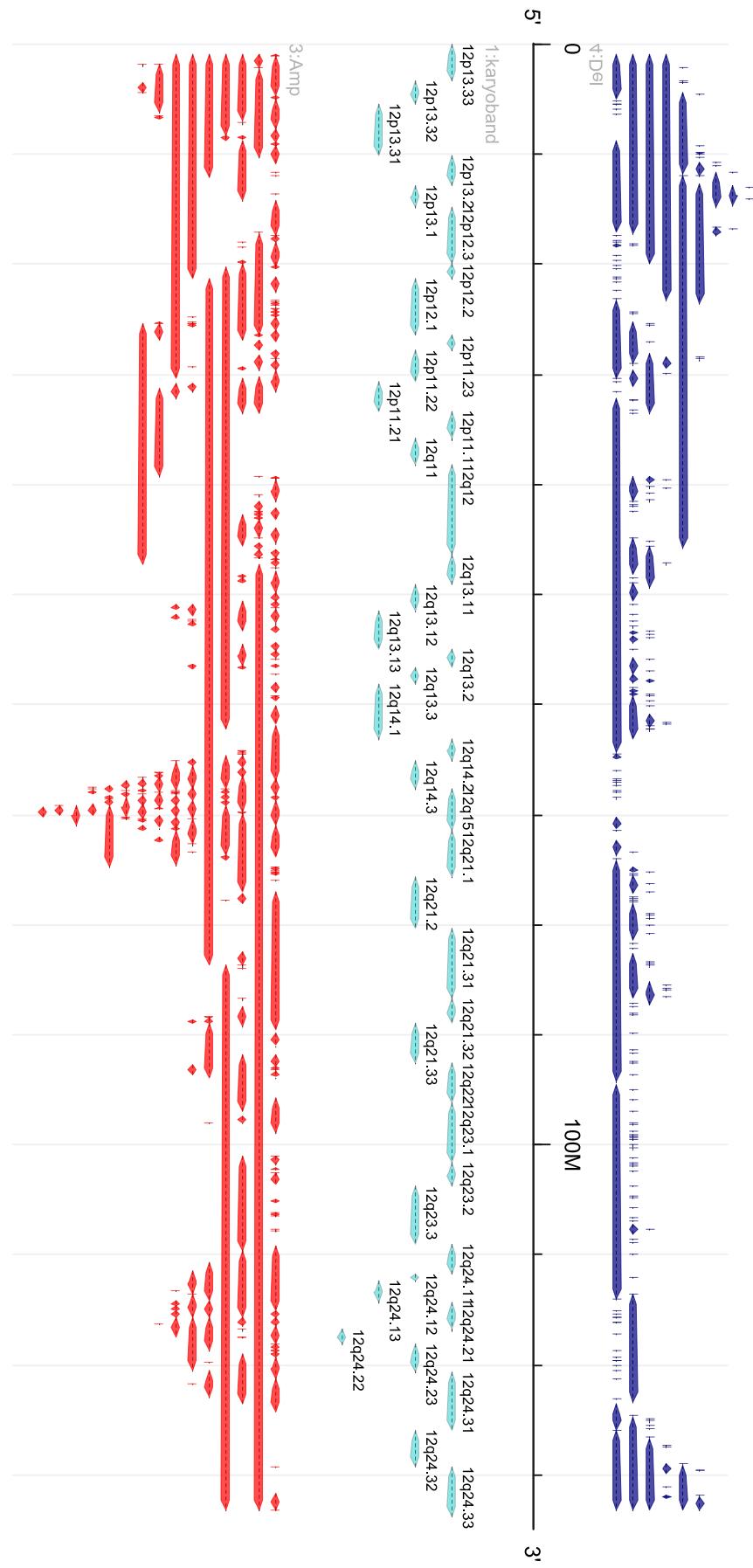


Abbildung A.16: Chromosom 12, BLCA, BGW6

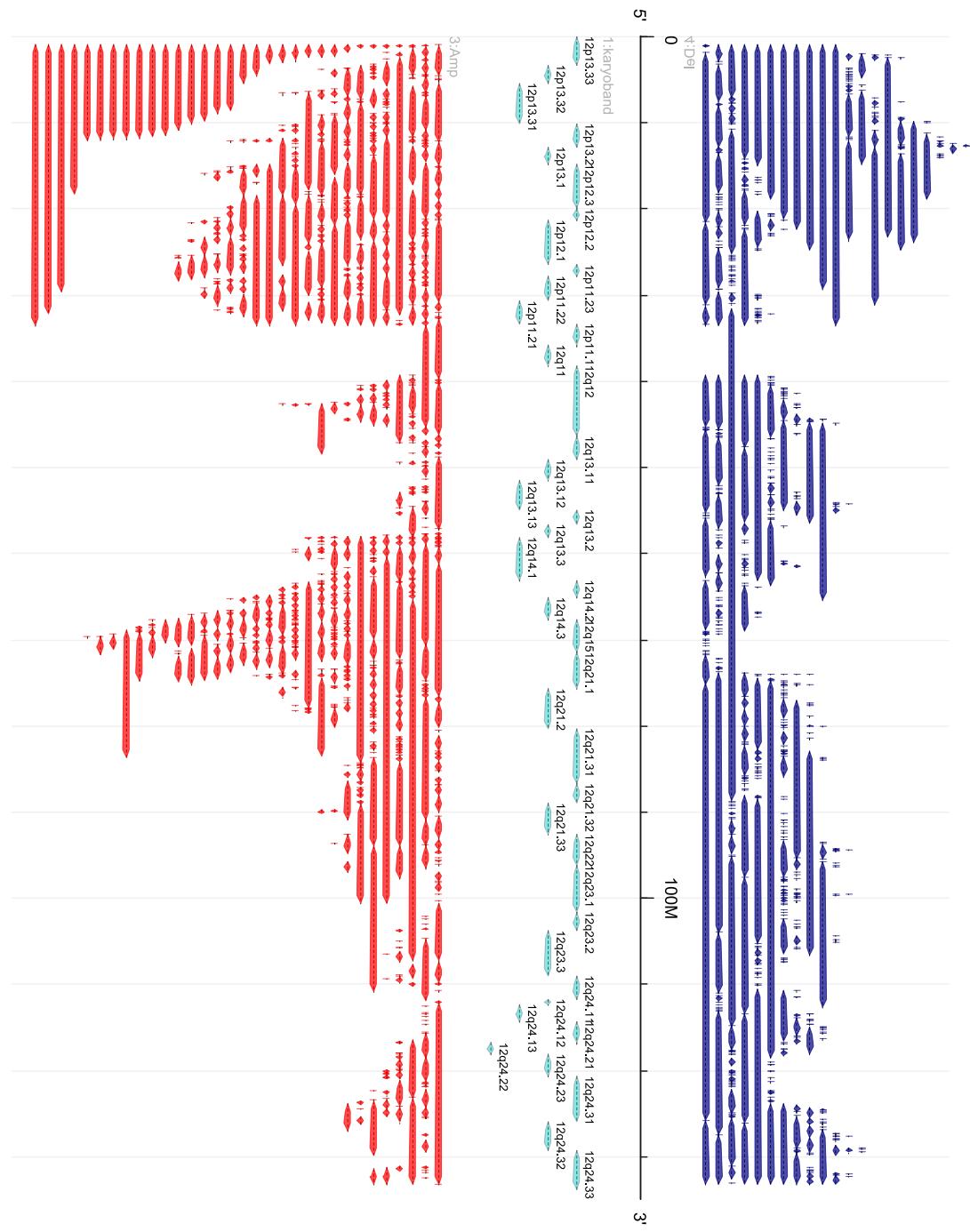


Abbildung A.17: Chromosom 12, BRCA, BGW6

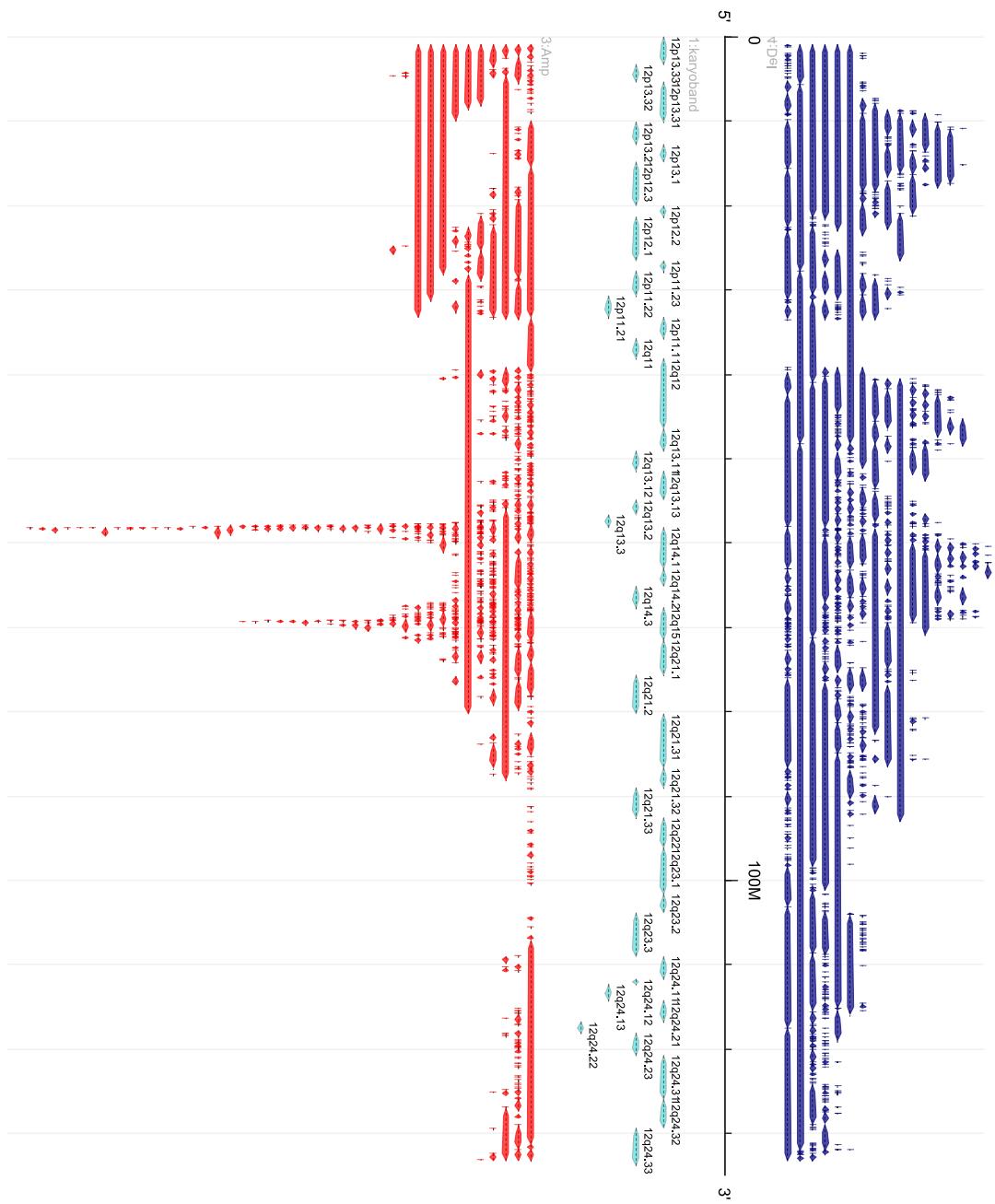


Abbildung A.18: Chromosom 12, GBM, BGW6

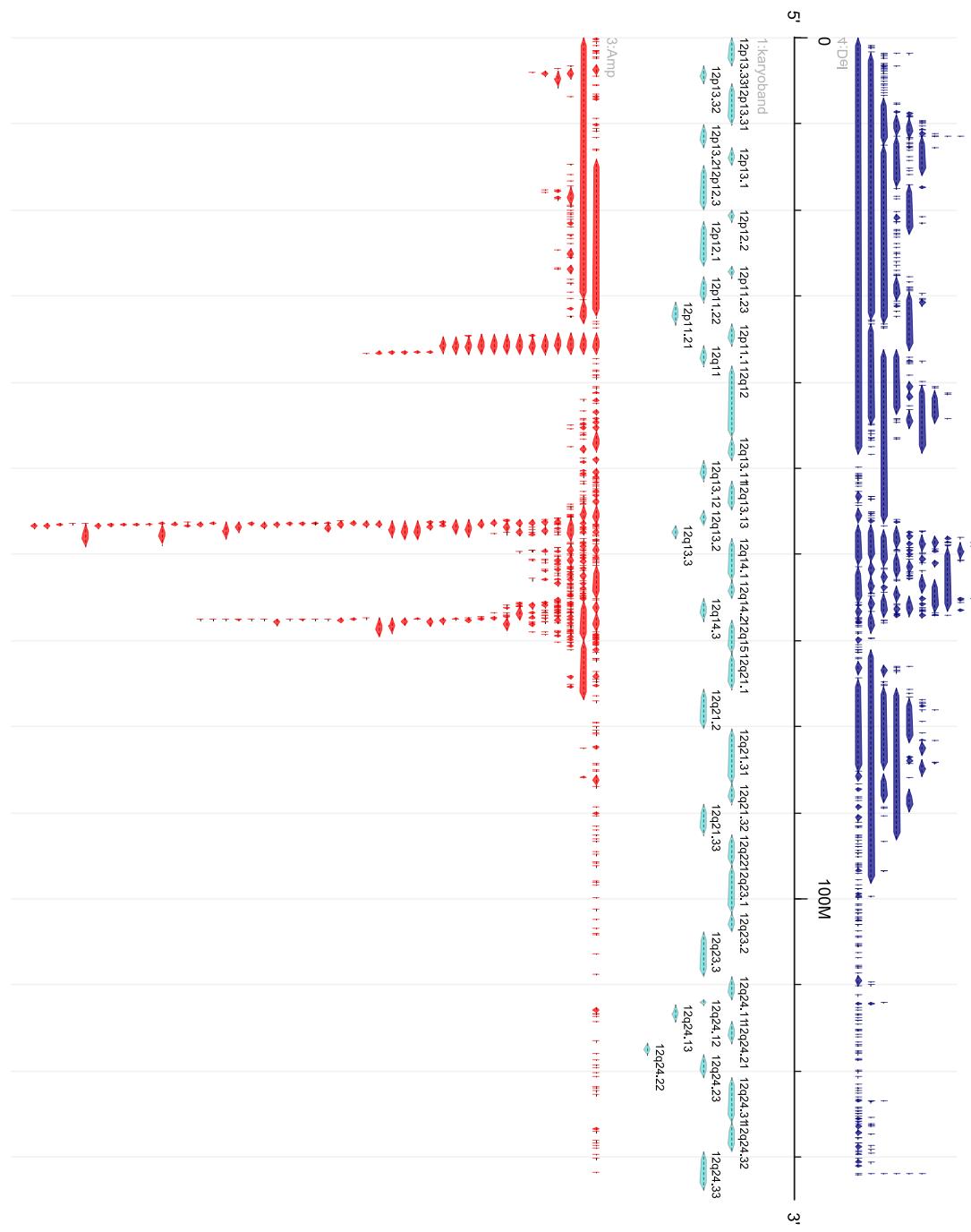


Abbildung A.19: Chromosom 12, GBM, HHH5

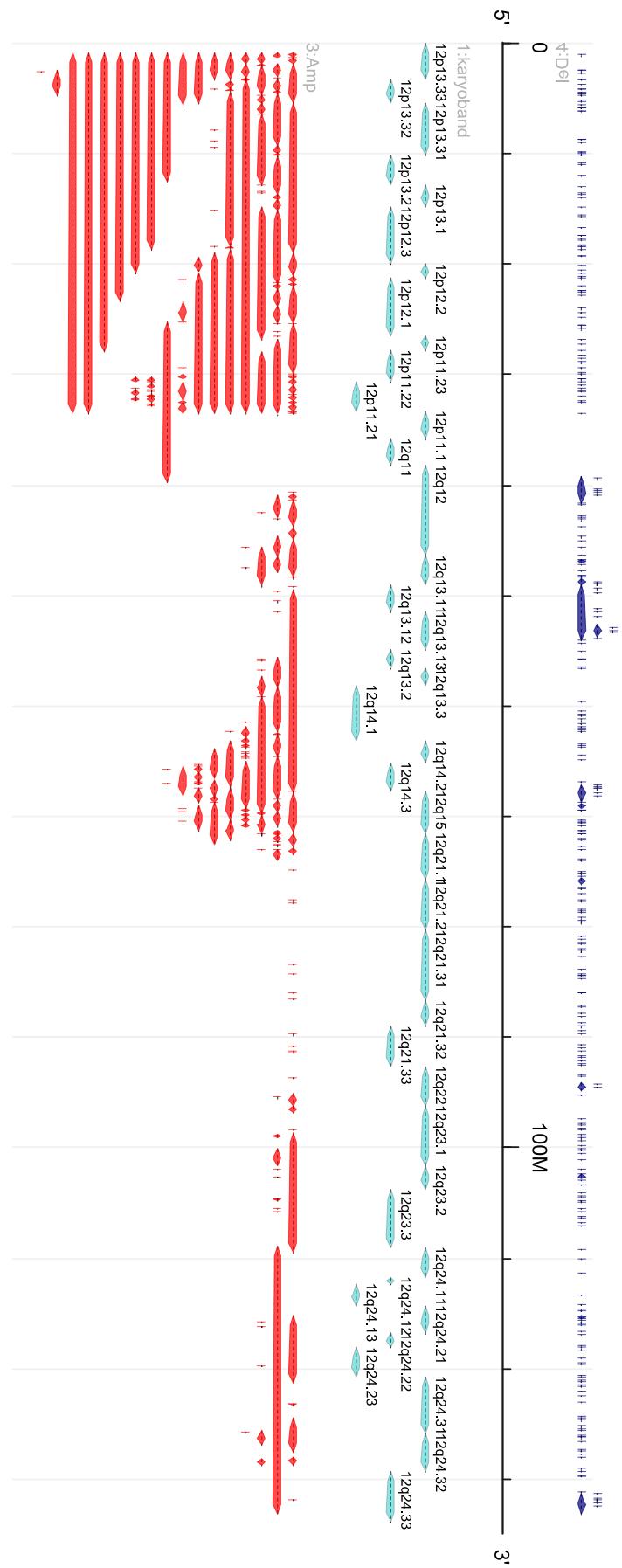


Abbildung A.20: Chromosom 12, HNSC, BGW6

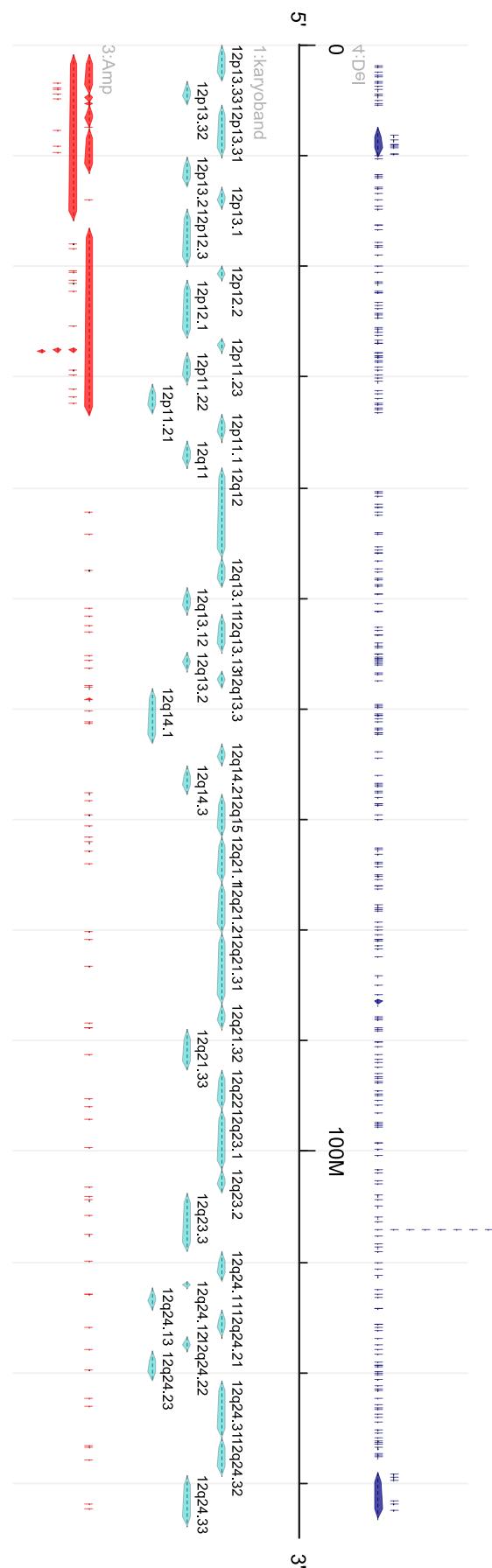


Abbildung A.21: Chromosom 12, KIRC, BGW6

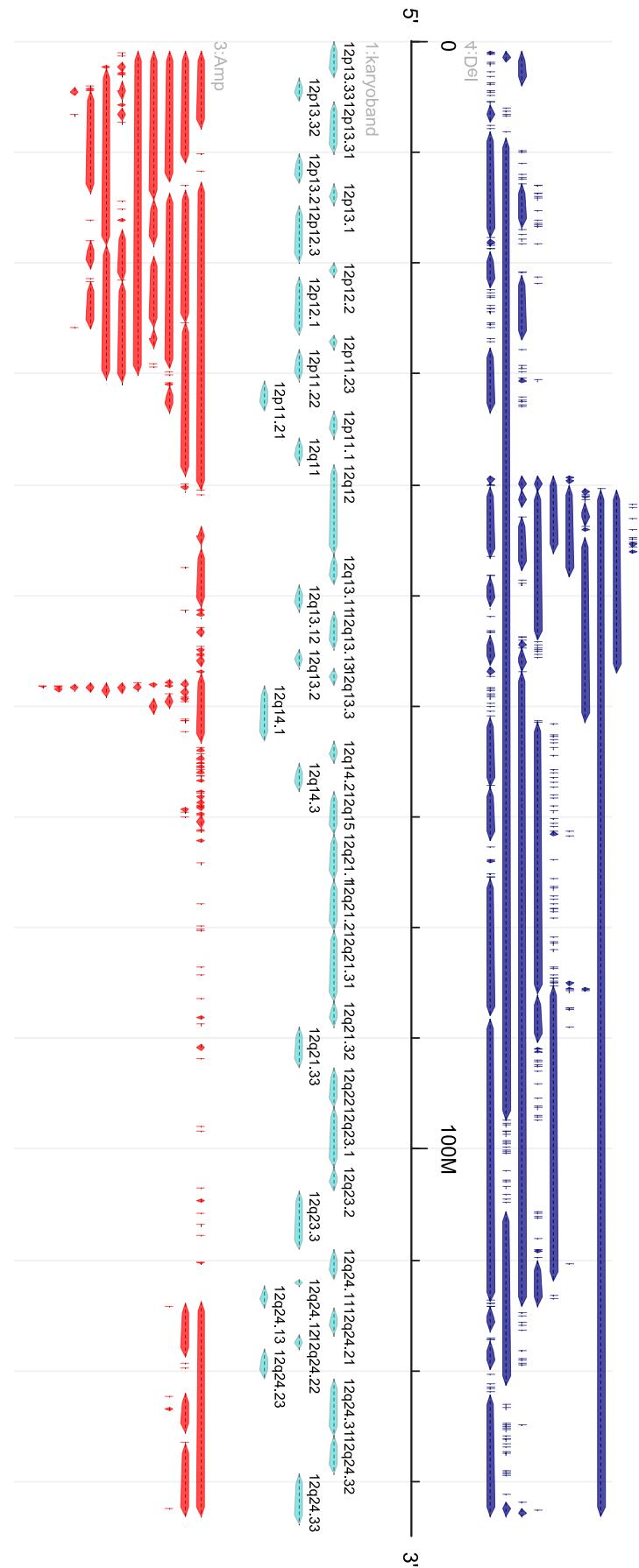


Abbildung A.22: Chromosom 12, LGG, BGW6

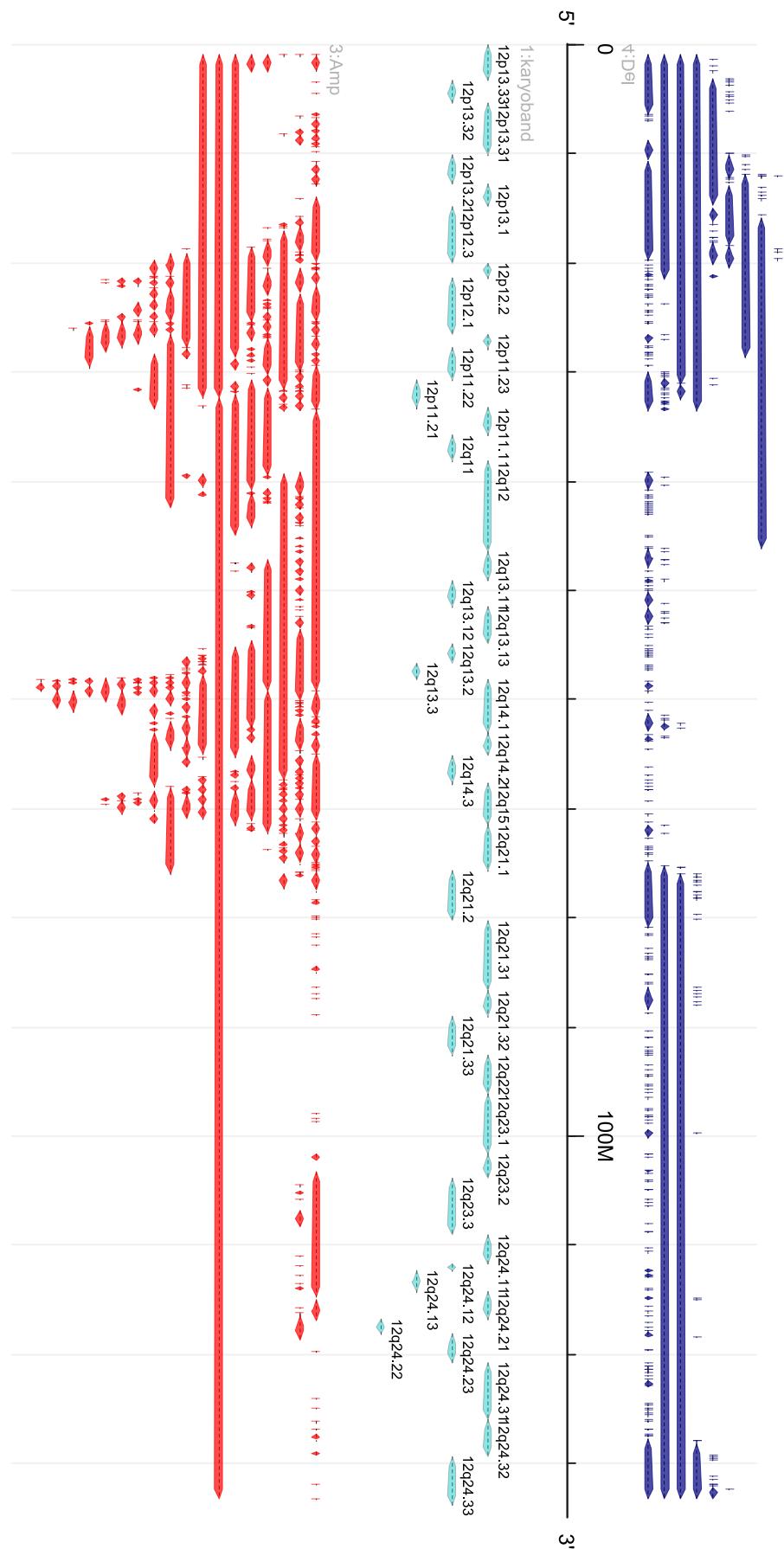


Abbildung A.23: Chromosom 12, LUAD, BGW6

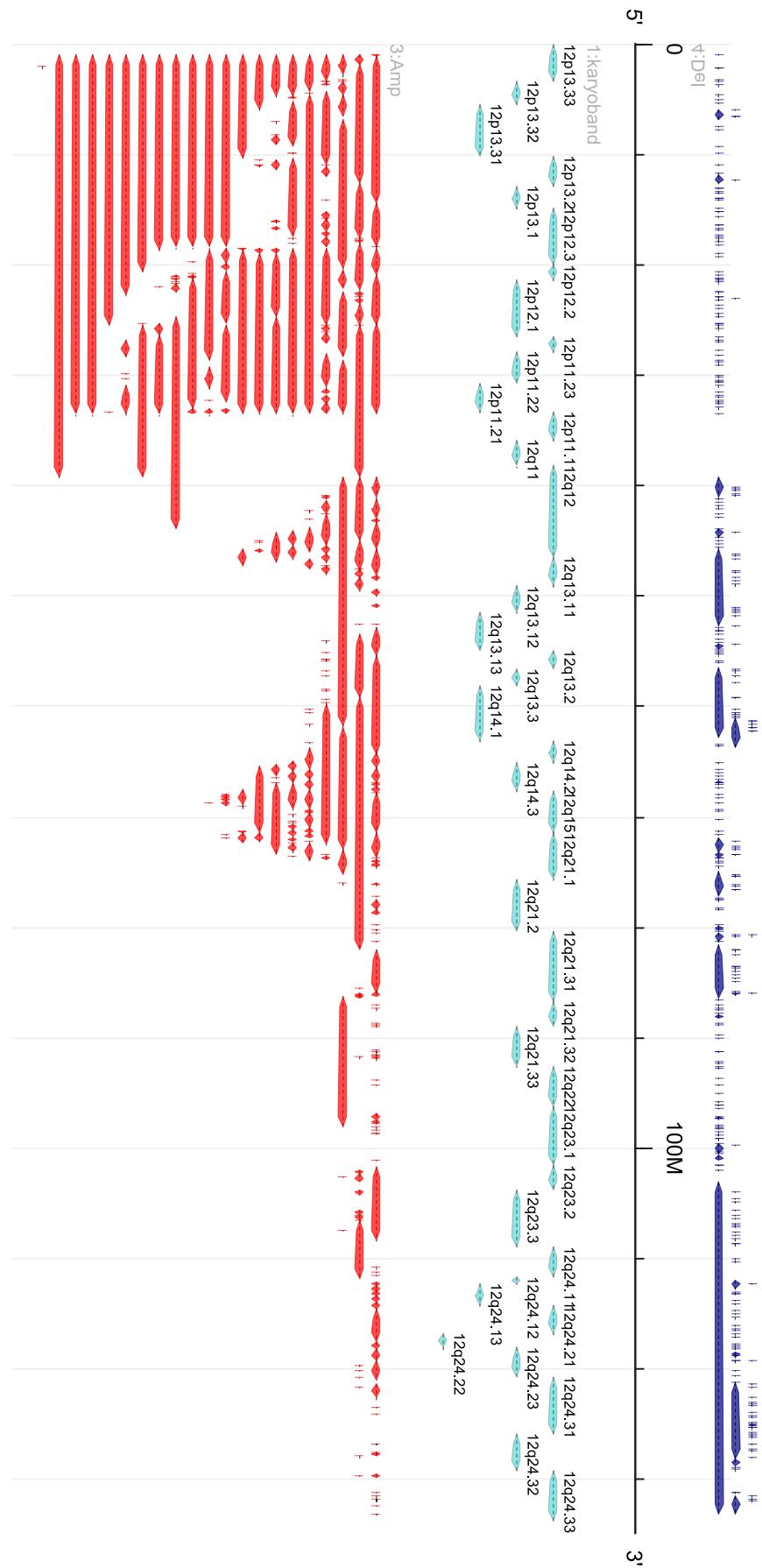


Abbildung A.24: Chromosom 12, LUSC, BGW6

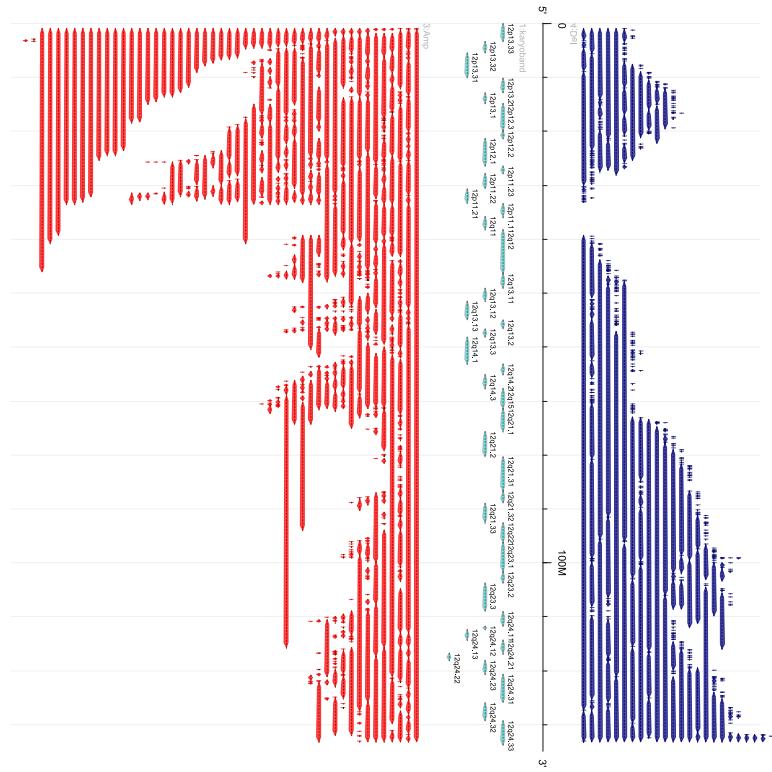


Abbildung A.25: Chromosom 12, OV, BGW6

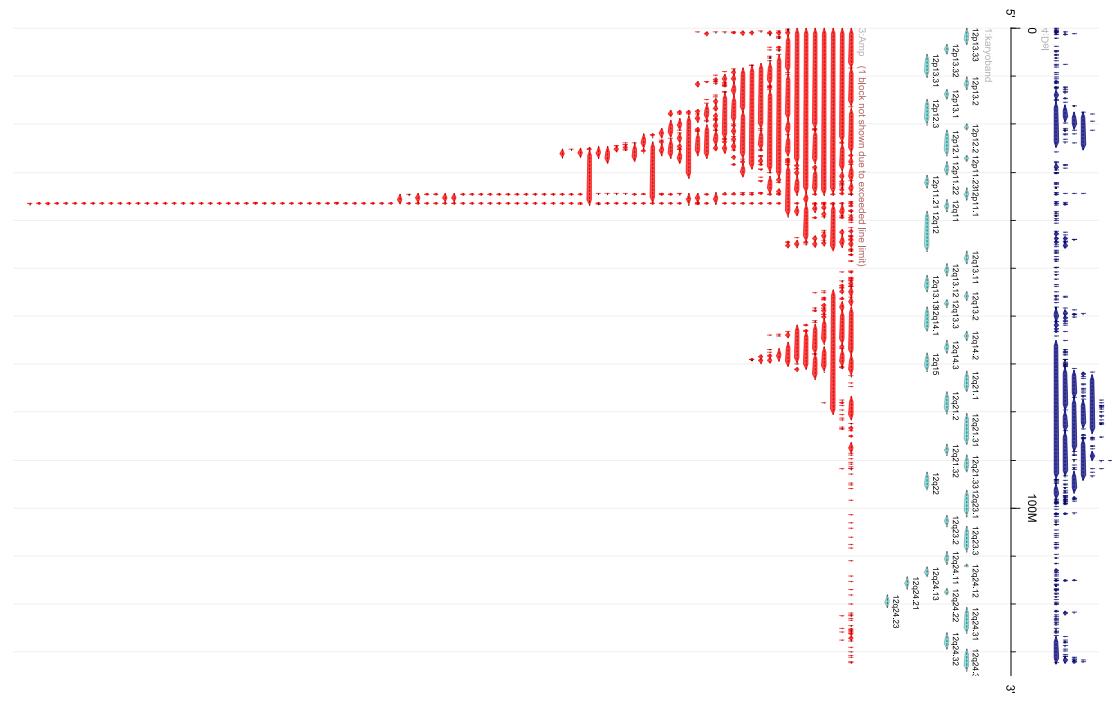


Abbildung A.26: Chromosom 12, OV, HH1M

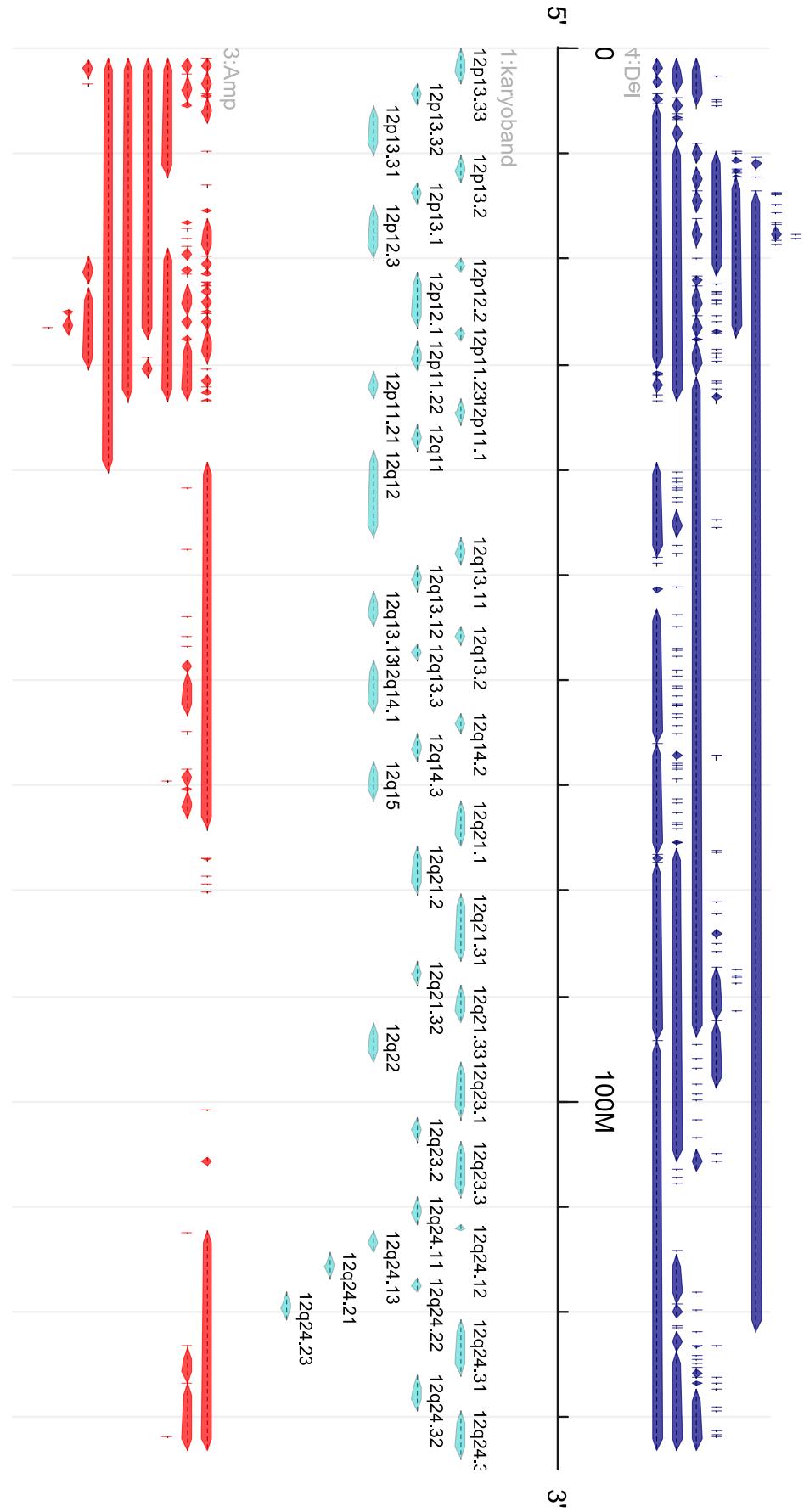


Abbildung A.27: Chromosom 12, READ, BGW6

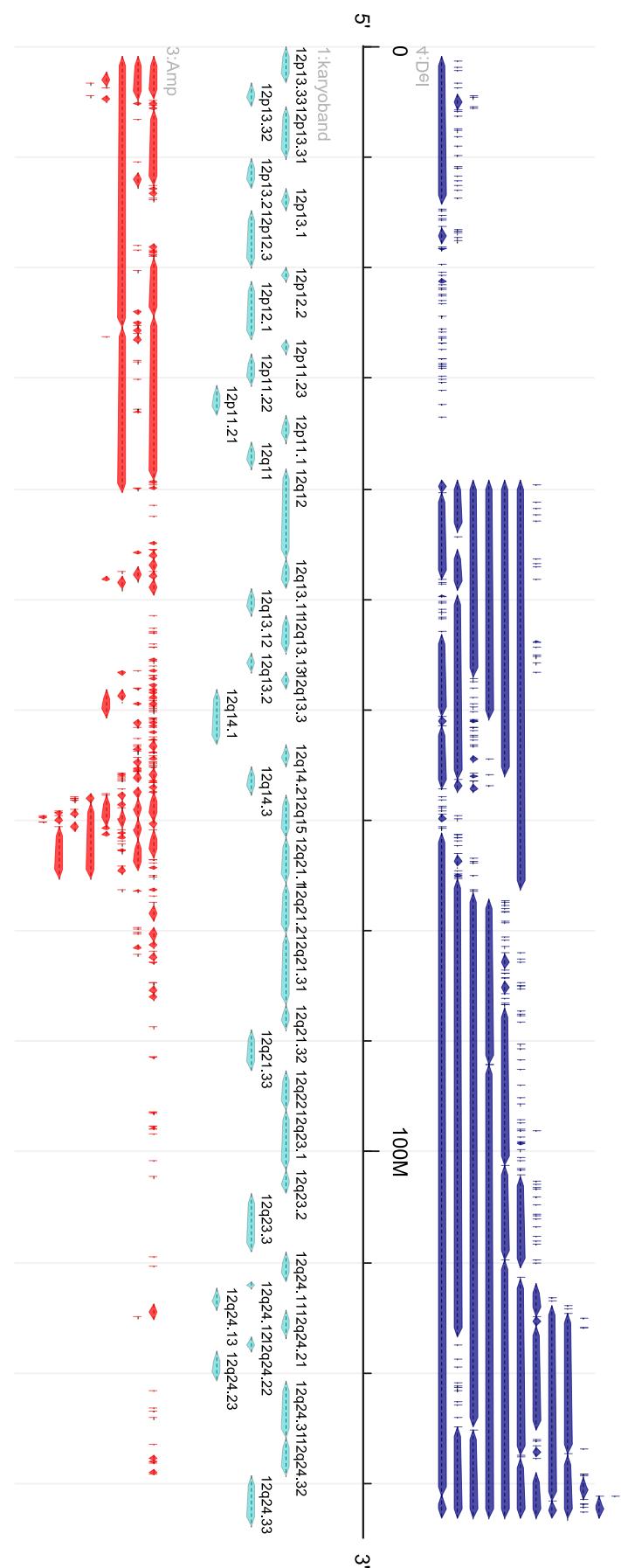


Abbildung A.28: Chromosom 12, SKCM, BGW6

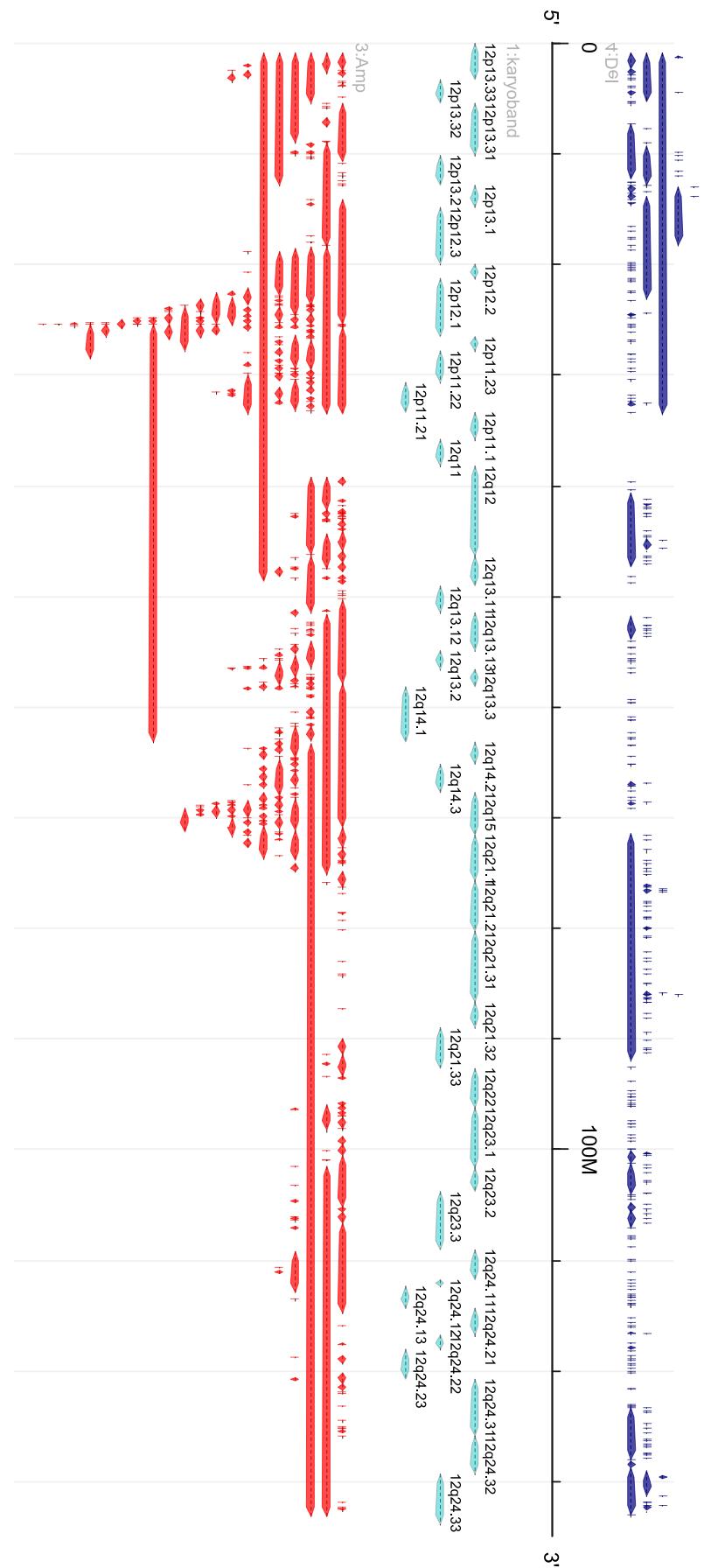


Abbildung A.29: Chromosom 12, STAD, BGW6

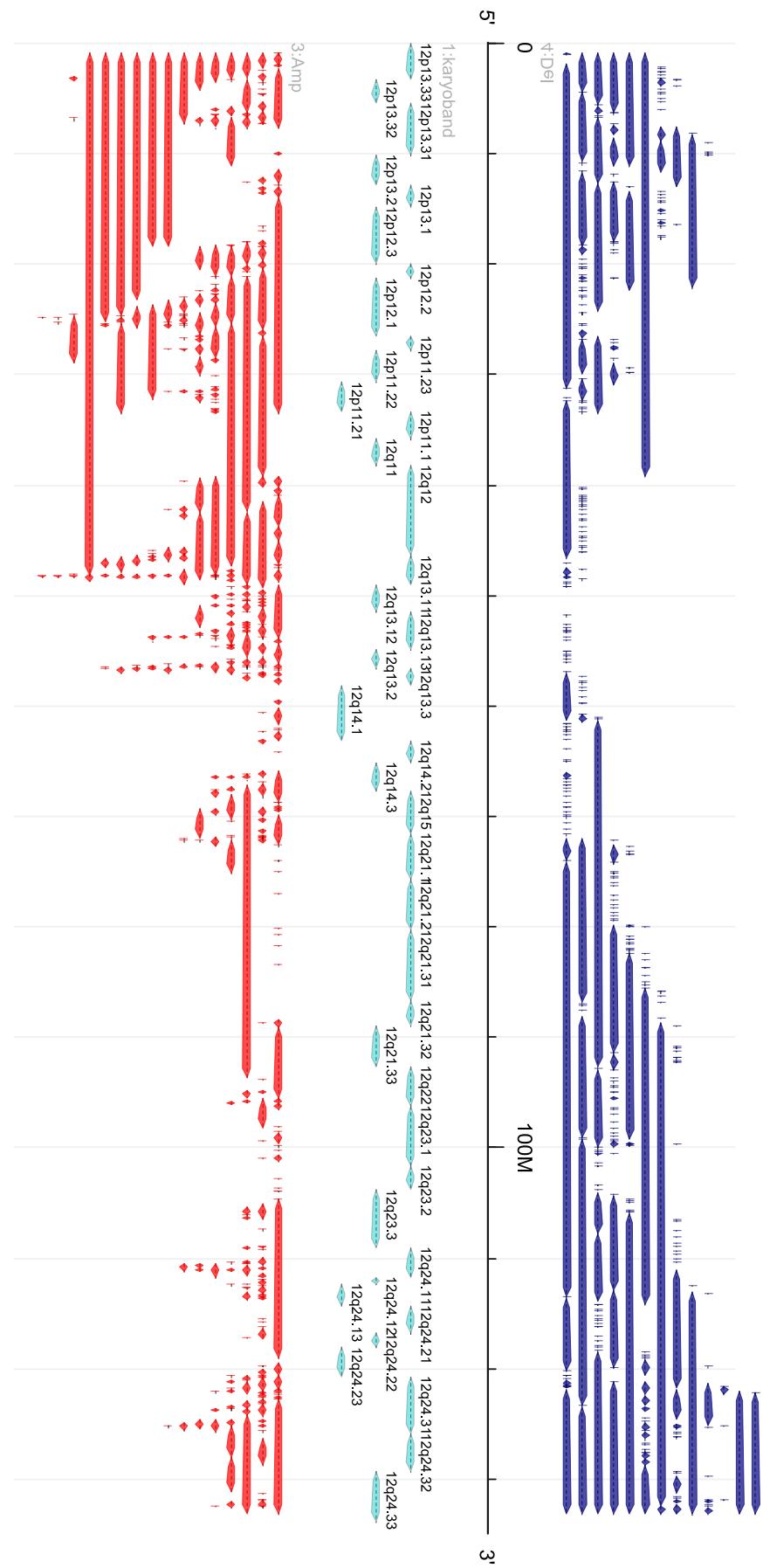


Abbildung A.30: Chromosom 12, UCEC, BGW6

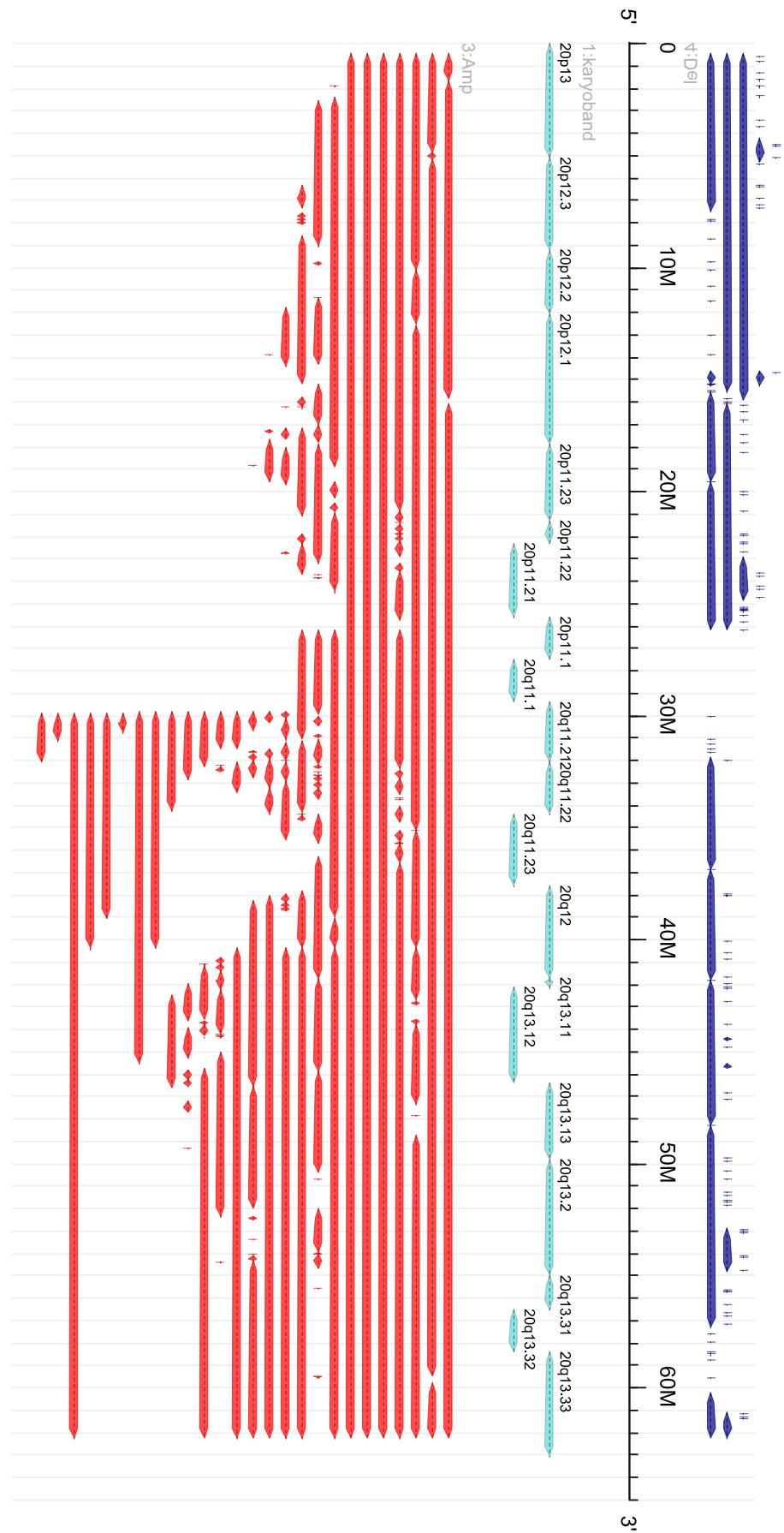


Abbildung A.31: Chromosom 20, BLCA, BGW6

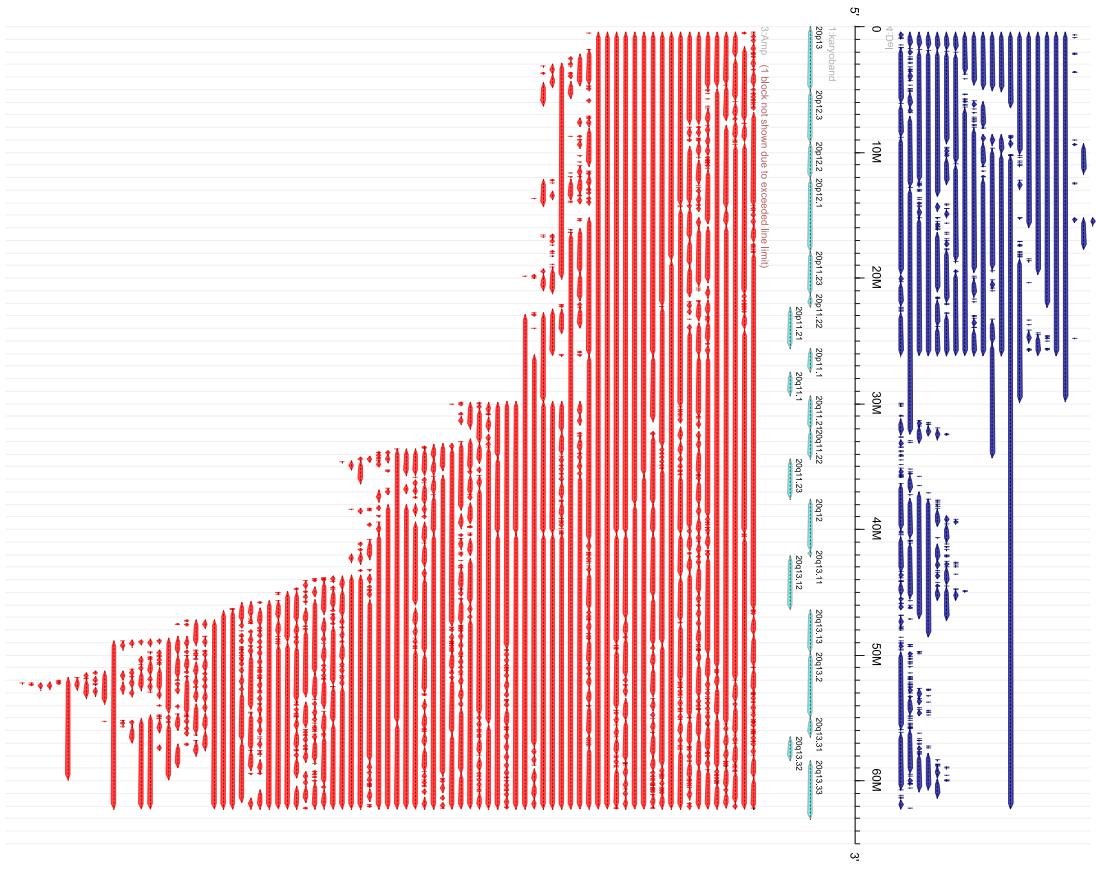


Abbildung A.32: Chromosom 20, BRCA, BGW6

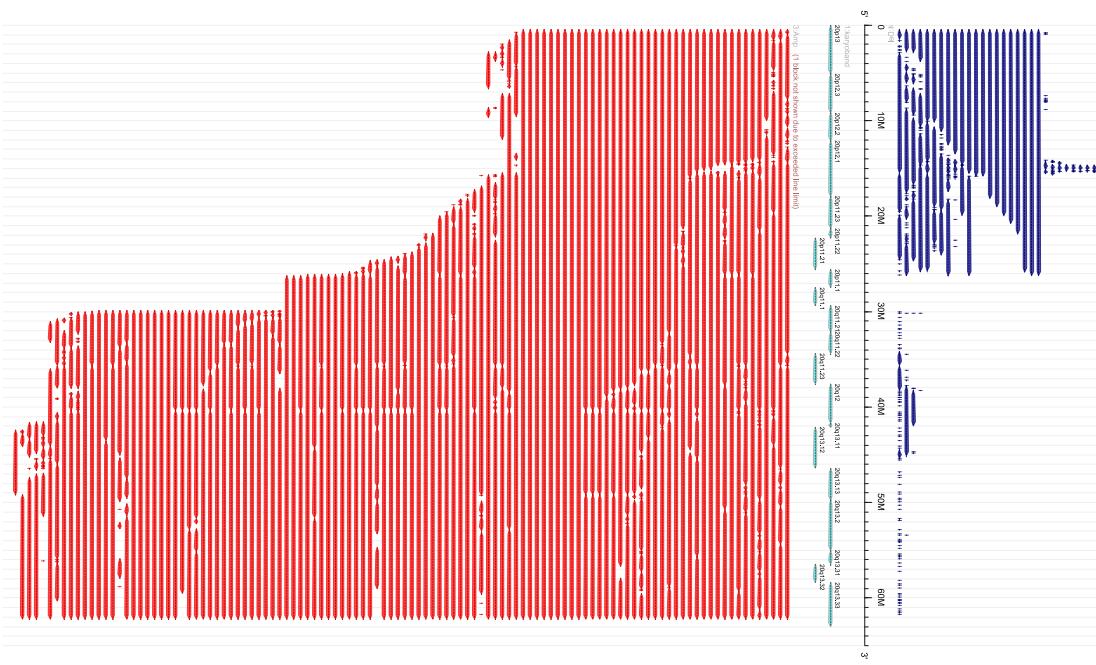


Abbildung A.33: Chromosom 20, COAD, BGW6

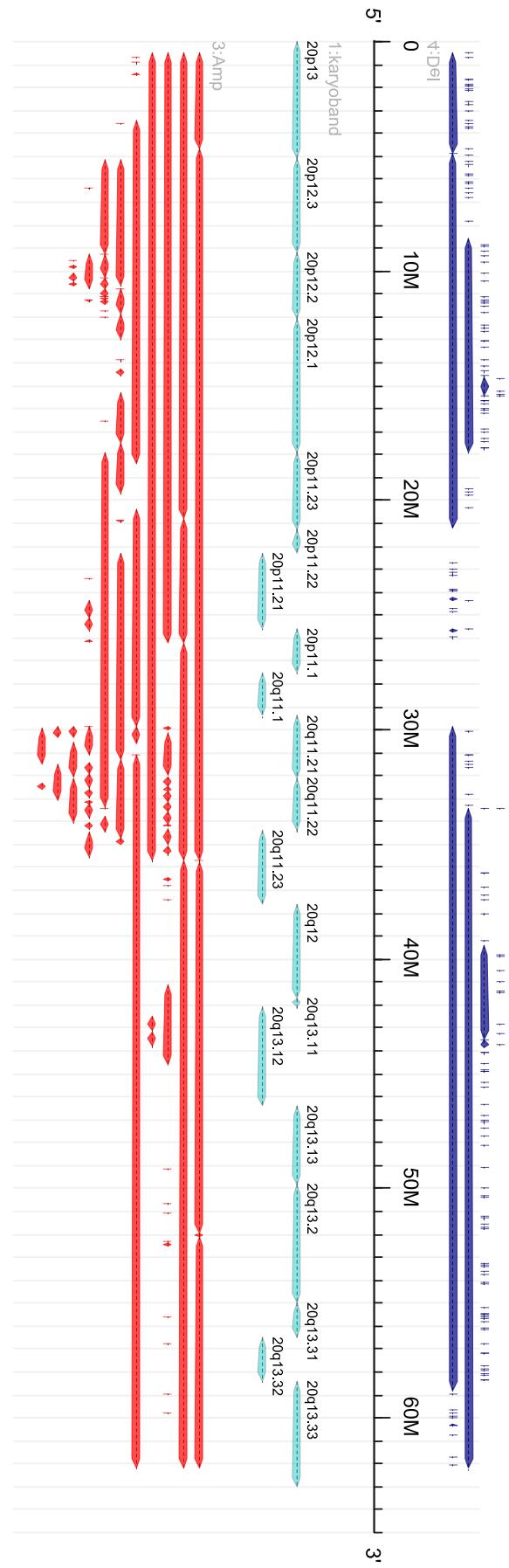


Abbildung A.34: Chromosom 20, HNSC, BGW6

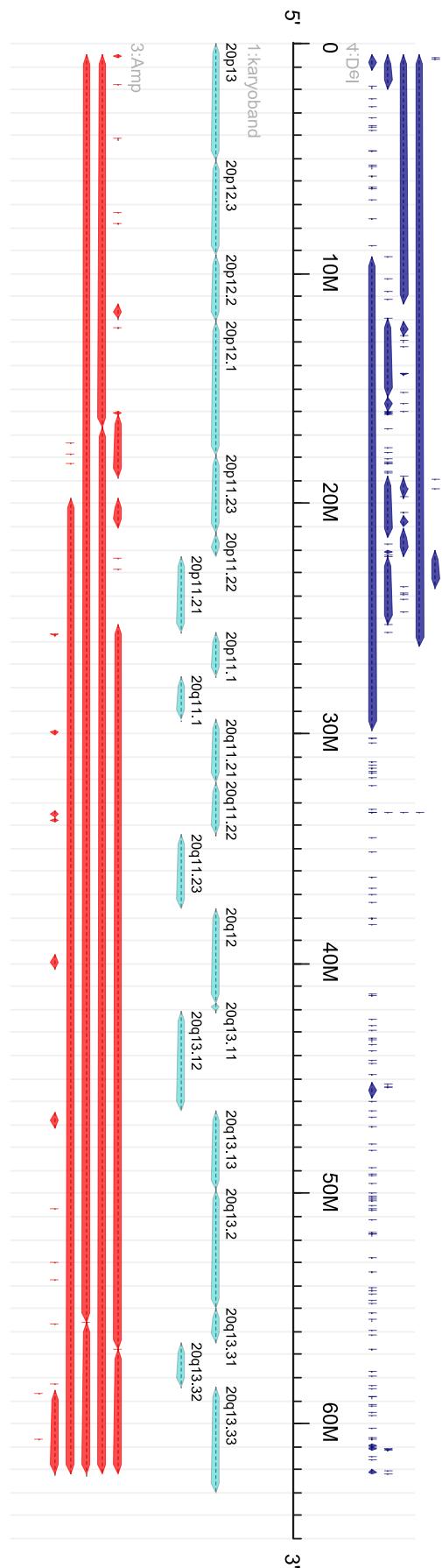


Abbildung A.35: Chromosom 20, LGG, BGW6

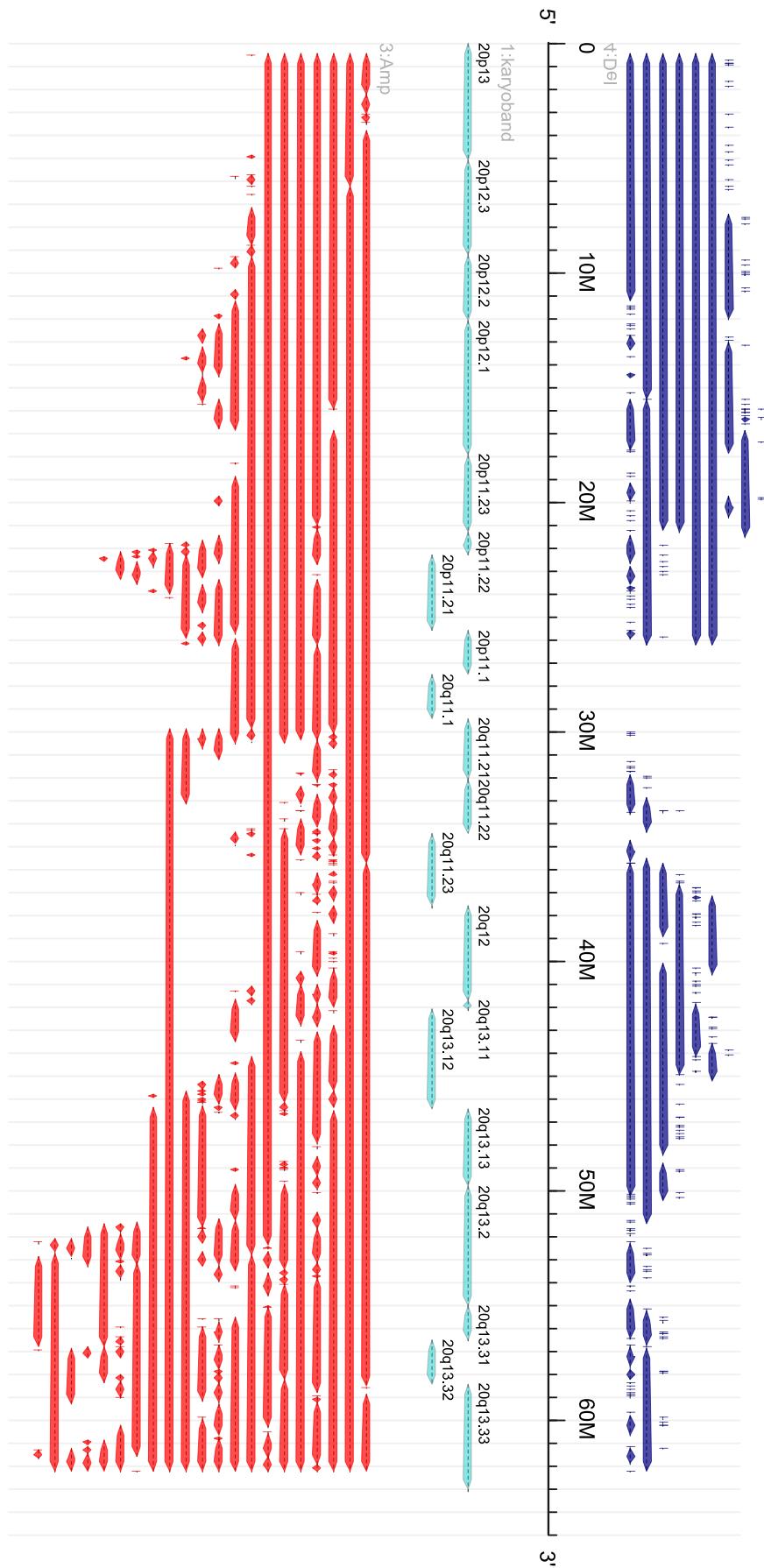


Abbildung A.36: Chromosom 20, LUAD, BGW6

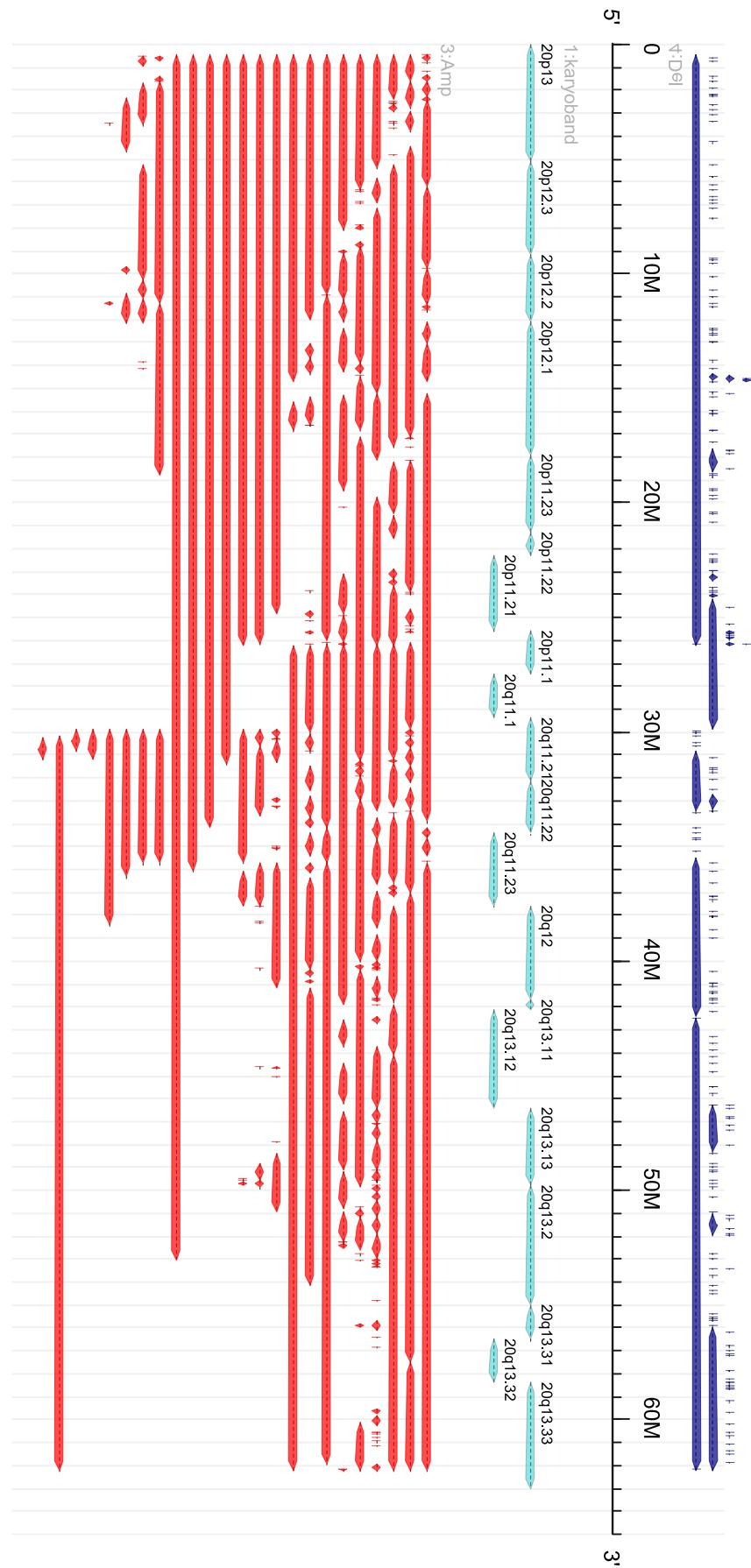


Abbildung A.37: Chromosom 20, LUSC, BGW6

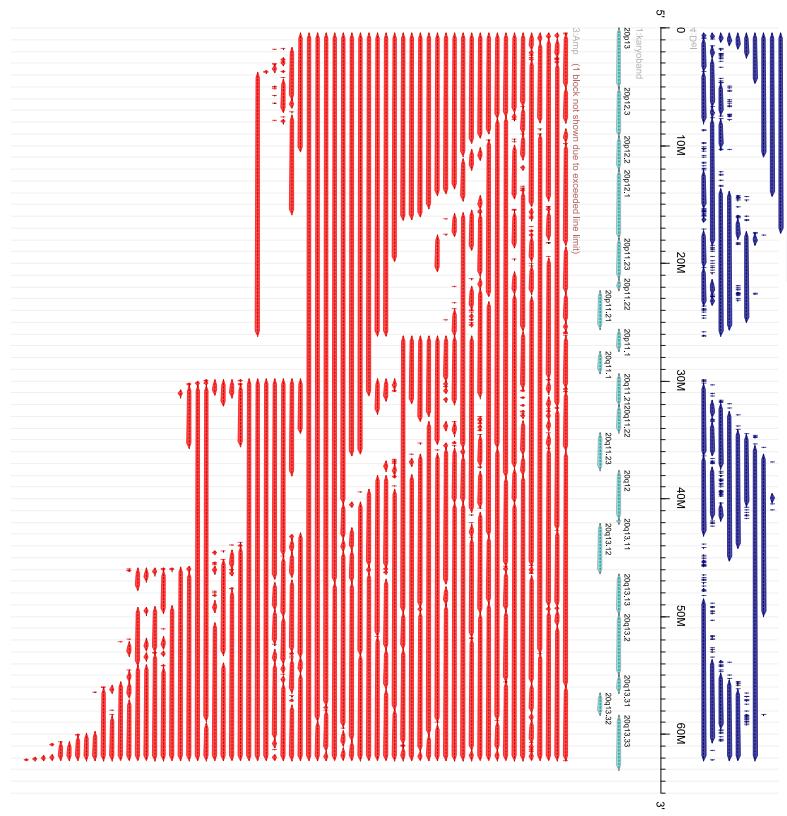


Abbildung A.38: Chromosom 20, OV, BGW6

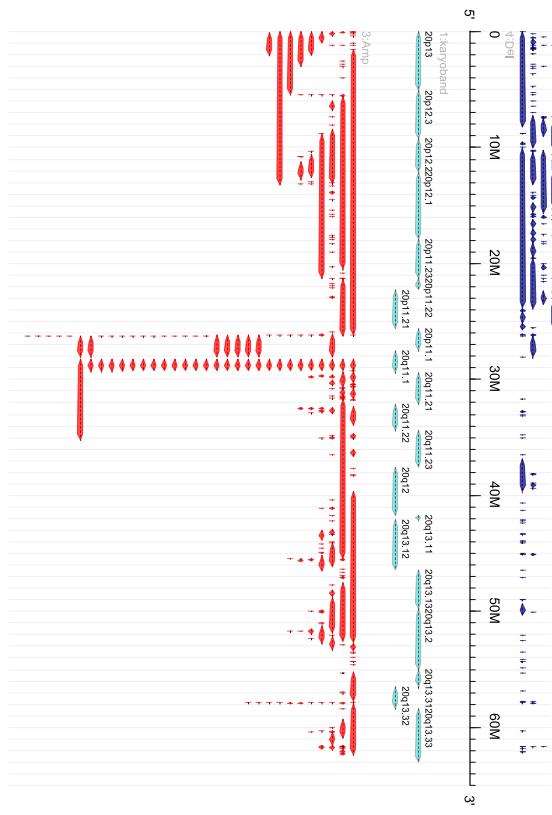


Abbildung A.39: Chromosom 20, OV, HH1M

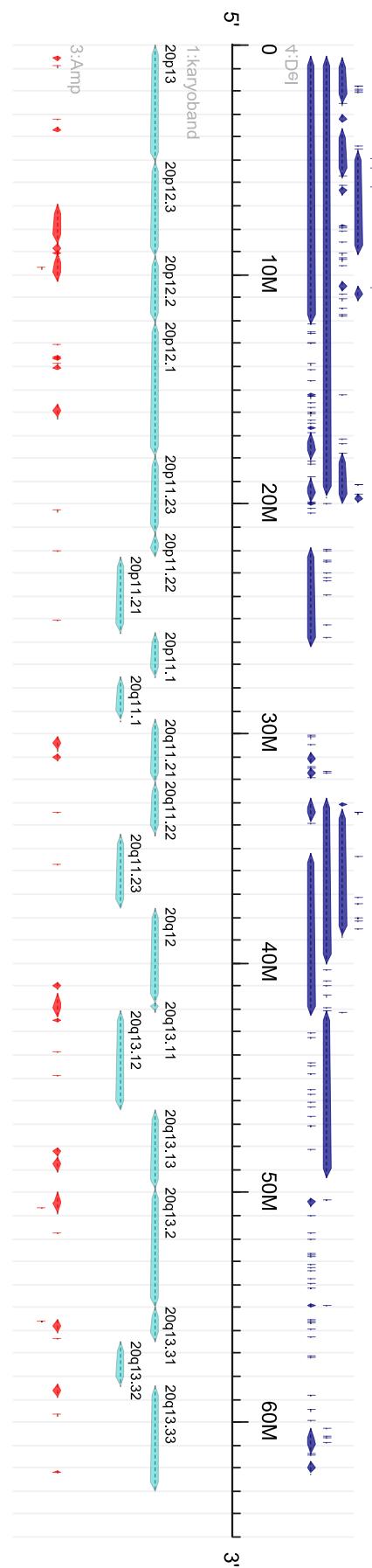


Abbildung A.40: Chromosom 20, PRAD, BGW6

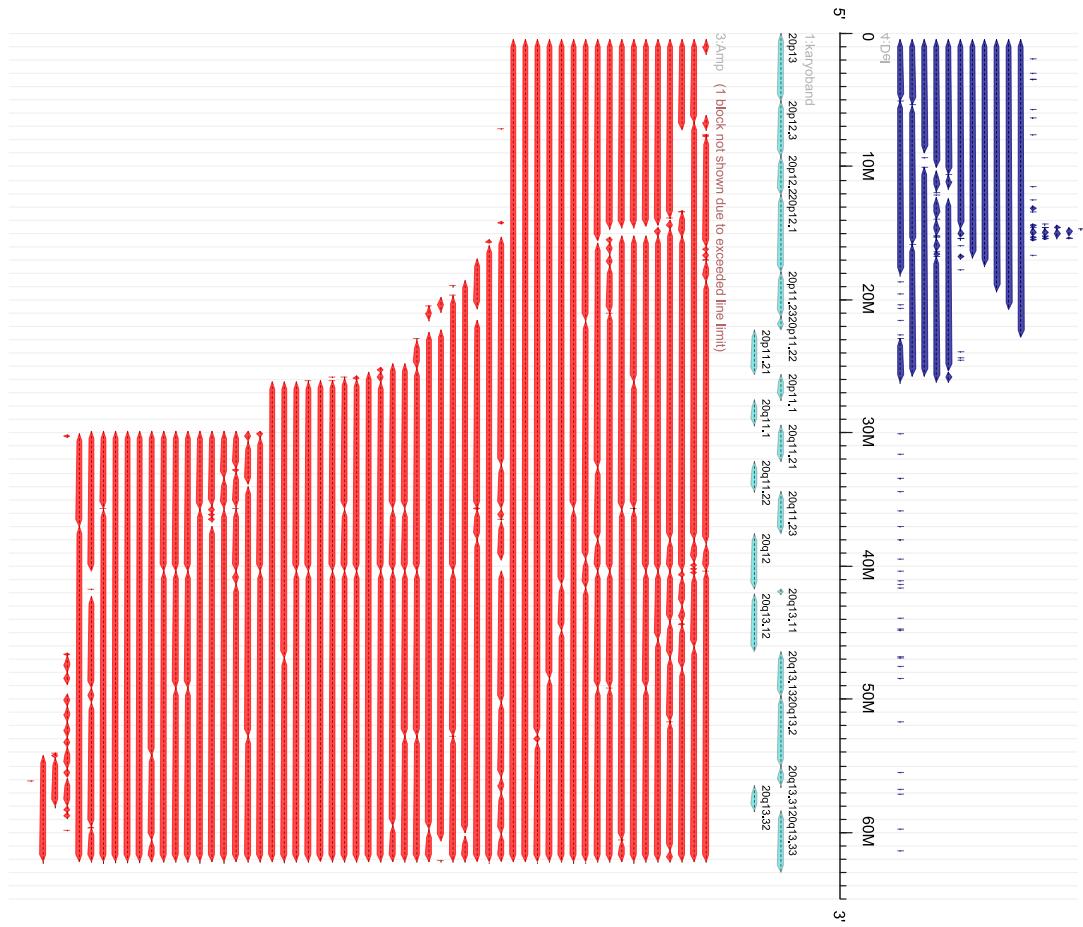


Abbildung A.41: Chromosom 20, READ, BGW6

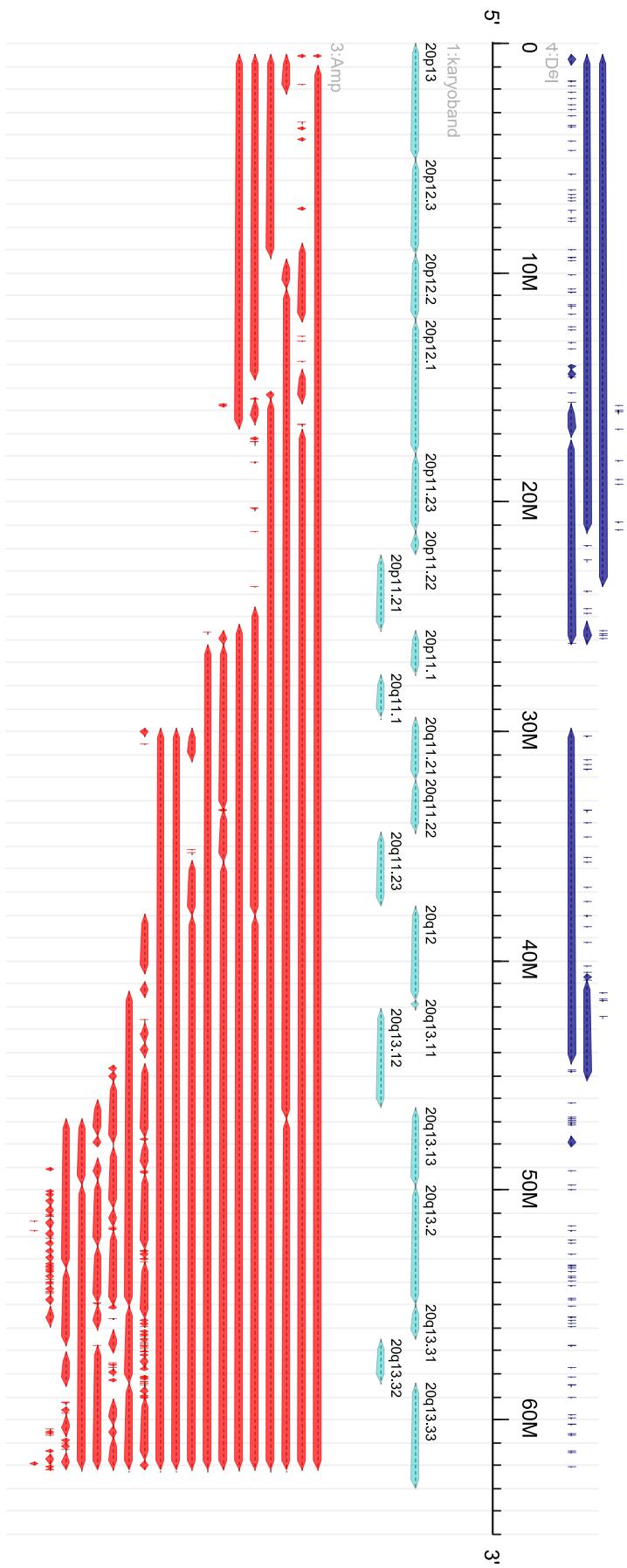


Abbildung A.42: Chromosom 20, SKCM, BGW6

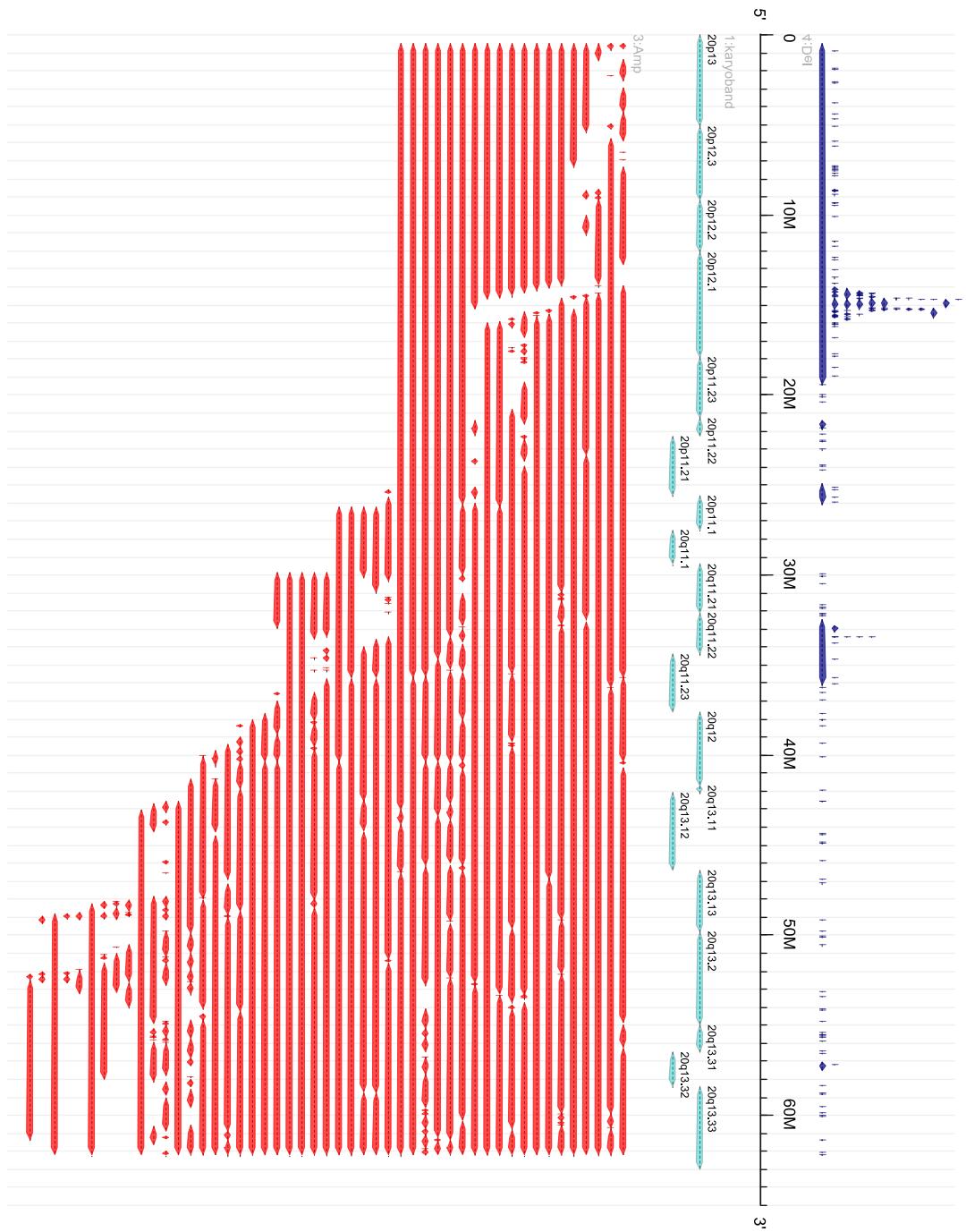


Abbildung A.43: Chromosom 20, STAD, BGW6

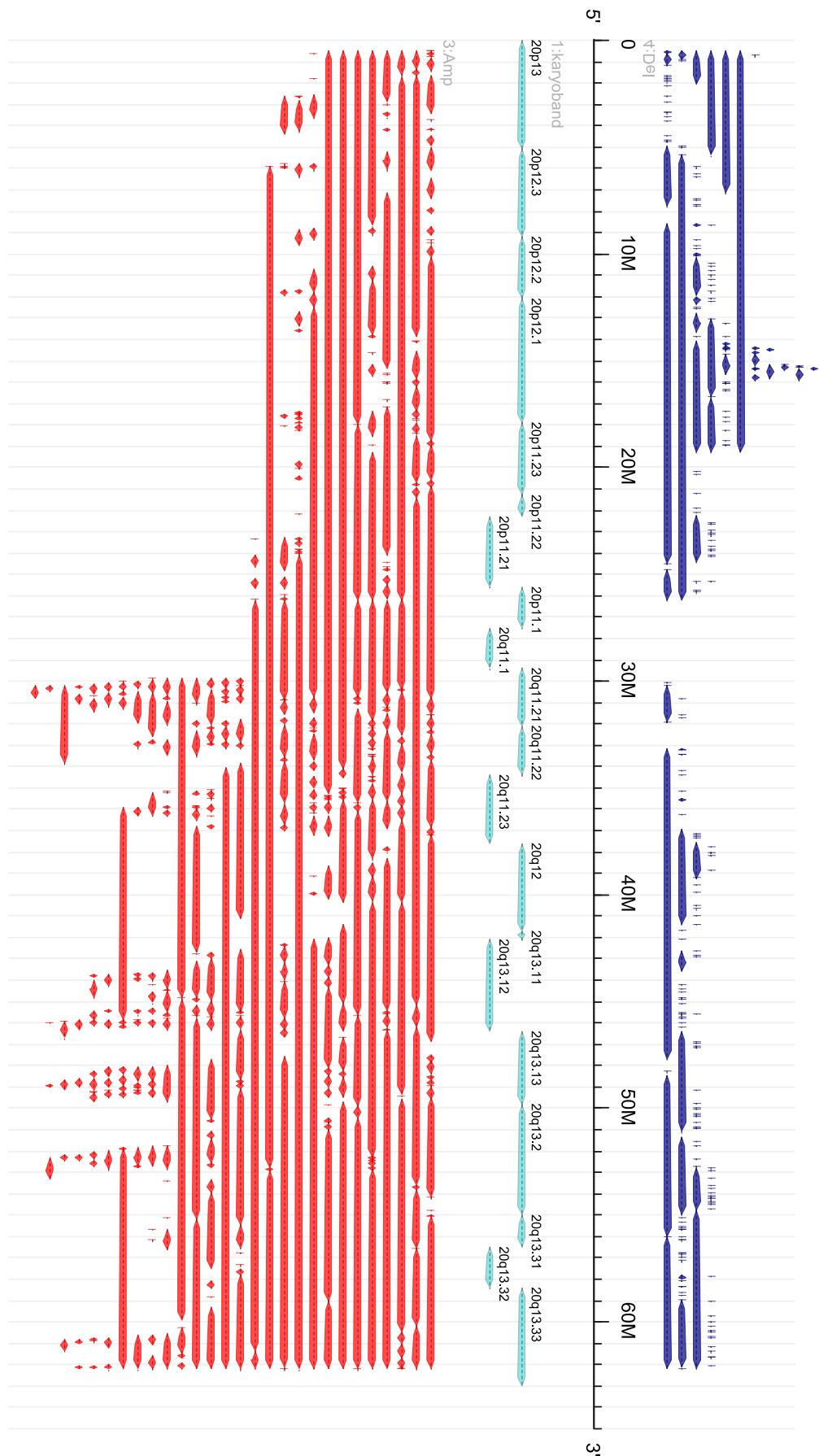


Abbildung A.44: Chromosom 20, UCEC, BGW6

Literaturverzeichnis

- [1] ISCN 2009. <http://atlasgeneticsoncology.org/ISCN09/ISCN09.html>.
 - [2] Affymetrix. <http://www.affymetrix.com/>.
 - [3] SNP array-based data TCGA National Cancer Institute Confluence Wiki. <https://wiki.nci.nih.gov/display/TCGA/SNP+array-based+data/#SNParray-baseddata-Level3Data>.
 - [4] The Genome Institute at Washington University. <http://genome.wustl.edu/>.
 - [5] TCGA barcode TCGA National Cancer Institute Confluence Wiki. <https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode#TCGAbBarcode-ReadingBarcodes>.
 - [6] HudsonAlpha. <http://hudsonalpha.org/>.
 - [7] Daniel C. Koboldt et al. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, September 2012.
 - [8] Veiko Krauß. Die struktur des zufalls als motor der veränderung. In *Gene, Zufall, Selektion*, pages 75–96. Springer, 2014.
 - [9] Malte Mader, Ronald Simon, and Stefan Kurtz. Fish oracle 2: a web server for integrative visualization of genomic data in cancer research. *Journal of Clinical Bioinformatics*, 4(5), March 2014.
 - [10] Malte Mader and Sascha Steinbiß. *The FISH Oracle user manual*. Zentrum für Bioinformatik, Universität Hamburg, Germany, March 2014.
 - [11] Craig Mermel, Steven Schumacher, Barbara Hill, Matthew Meyerson, Rameen Beroukhim, and Gad Getz. Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.
 - [12] Ryan E Mills, Klaudia Walter, Chip Stewart, Robert E Handsaker, Ken Chen, Can Alkan, Alexej Abyzov, Seungtai Chris Yoon, Kai Ye, R Keira Cheetham, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011.
-

- [13] National Library of Medicine (US). Genetics Home Reference. <http://ghr.nlm.nih.gov/>.
- [14] The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322, January 2014.
- [15] Broad Institute of MIT and Harvard. <http://www.broadinstitute.org/>.
- [16] Fish Oracle. <http://fishoracle.zbh.uni-hamburg.de:8080/fishoracle/>.
- [17] The Cancer Genome Atlas Data Portal. <https://tcga-data.nci.nih.gov/tcga/tcgahome2.jsp>.
- [18] Gordon Saksena, Barbara Tabak, and Jeff Gentry. *Generates segmented copy number calls from raw tumor and normal SNP6 CEL files*. MIT and Harvard, 2013.
- [19] Illumina | Sequencing and array-based solutions for genetic research. <http://www.illumina.com/>.
- [20] Weizmann Institute of Science. The Human Gene Compendium. <http://www.genecards.org/>.

Eidesstattliche Erklärung

Ich versichere, dass ich die vorstehende Arbeit selbstständig und ohne fremde Hilfe angefertigt und mich anderer als der im beigefügten Verzeichnis angegebenen Hilfsmittel nicht bedient habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht.

Sebastian Vincent Weigel